

CORNERSTONE COMMUNITY COLLEGE

BALBINO SOUZA RAMOS NETO

CARDIOVASCULAR DISEASE DATASET ANALYSIS

DS 204 - Analysis for Data Science

Vancouver - BC / Canada

2026

BALBINO SOUZA RAMOS NETO

CARDIOVASCULAR DISEASE DATASET ANALYSIS

DS 204 - Analysis for Data Science

Final project, presented to Cornerstone Community College, as part of the requirements to obtain a grade in the course “DS 204” for Data Science program.

Vancouver, January 30 of 2026.

TUTOR

Prof. M.Eng. Atabak Eghbal

Glossary

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | SBP |
6. Blood Pressure |BP|
7. Diastolic blood pressure | Examination Feature | ap_lo | DBP |
8. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
9. Presence or absence of cardiovascular disease | cardio | binary |
10. Cardiovascular Disease |CVD|
11. Body Mass Index |BMI|
12. World Health Organization |WHO|
13. Odds Ratio |OR|
14. Risk Difference |RD|
15. Risk Ratio |RR|
16. 95% confidence interval |95% CI|
17. Analysis of Variance |ANOVA|

Introduction

Cardiovascular disease (CVD) is an umbrella term covering coronary artery disease, cerebrovascular disease, peripheral arterial disease, and other atherosclerotic conditions that remain leading contributors to morbidity and mortality, with risk shaped by both non-modifiable factors (age, sex) and modifiable cardiometabolic exposures (blood pressure, adiposity, lipids, glucose status, and behaviors). Global evidence indicates that modifiable factors—including hypertension, smoking, diabetes, abdominal obesity, diet, and physical activity—account for a large proportion of first myocardial infarction risk (INTERHEART), motivating analytics that quantify how these factors co-occur and relate to CVD markers in clinical datasets.

This project analyzes a large clinical-examination dataset (cardiotrain.csv, $n = 70,000$) containing objective measures (age, height, weight, gender), examination measures (systolic/diastolic blood pressure, cholesterol category, glucose category), and lifestyle behaviors (smoking, alcohol intake, physical activity), with a binary target indicating presence/absence of CVD (cardio). The project goal is to quantify association strength (not causality) between CVD status and (i) systolic blood pressure, (ii) BMI, and (iii) their joint presence, and to provide interpretable public-health insights.

Methods

Feature engineering used standard clinical definitions (according to WHO definitions). Age was converted from days to years as **AgeYears** = $\frac{Age\ Days}{365.25}$, and BMI was computed as **BMI** = $\frac{Weight\ (kg)}{Height\ (m)^2}$. For blood pressure analyses, a physiological plausibility filter was applied (ap_hi 70–250, ap_lo 40–150, and $ap_hi \geq ap_lo$) to reduce measurement artifacts; 68,668 of the 70,000 records passed this filter.

Hypothesis testing used Welch's two-sample t-test for comparing group means when variances may differ, where $t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_1^2/n_1 + s_0^2/n_0}}$, and p-values were computed under the corresponding Welch degrees of freedom. For 3-group comparisons (cholesterol categories 1/2/3), one-way ANOVA was used, with $F = \frac{MS_{between}}{MS_{within}}$, and effect size reported using $\eta^2 = \frac{SS_{between}}{SS_{total}}$. To compare "toxicity/impact" strictly as association strength, we reported standardized differences and risk measures rather than causal language. Standardized mean difference was computed as Cohen's $d = (\bar{x}_1 - \bar{x}_0)/S_p$ (with pooled SD S_p), and for binary outcomes we reported risk ratio $RR = \frac{P(Y=1|E=1)}{P(Y=1|E=0)}$ and odds ratio $OR = \frac{a/b}{c/d}$ from 2x2 tables.

For correlation test, Pearson r (linear association), Spearman ρ (rank monotonic association), Kendall T (rank concordance), point-biserial r_{pb} (binary–continuous), and Phi ϕ (binary–binary).

Derived variables (used to increase clinical interpretability):

AgeYears = AgeDays / 365.25.

Height_m = Height_cm / 100.

BMI = Weight_kg / (Height_m²).

Hypertension (binary): hypertensive = 1 if (SBP ≥ 140) OR (DBP ≥ 90), computed only on BP-plausible records.

Obesity (binary): obese = 1 if BMI ≥ 30.

Why "association strength" and not causality: all measurements are observed at a single examination timepoint, so temporal direction (cause → effect) cannot be established using this dataset.

Exploratory data analysis (EDA)

The dataset contains 70,000 examinations and an overall observed CVD prevalence of 49.97%. Continuous-variable descriptive statistics are shown below.

Variable	N	Mean	Median	Std. dev.
Age (years)	70,000	53.3029	53.9439	6.755
Height (cm)	70,000	164.3592	165	8.2101
Weight (kg)	70,000	74.2057	72	14.3958
BMI (kg/m ²)	70,000	27.5565	26.3741	6.0915
Systolic BP ap_hi (mmHg, plausible only)	68,668	126.6688	120	16.6811
Diastolic BP ap_lo (mmHg, plausible only)	68,668	81.302	80	9.4233

Hypothesis testing

Hypothesis 1 (BMI difference by CVD status):

Welch two-sample t-test (used for BMI and SBP differences by cardiac status):

H0: $\mu_1 \leq \mu_0$ versus H1: $\mu_1 > \mu_0$ (one-sided “greater” direction defined a priori).

Test statistic (conceptual): $t = (\text{mean1} - \text{mean0}) / \sqrt{s_1^2/n_1 + s_0^2/n_0}$.

We also report Cohen’s d as a standardized mean difference: $d = (\text{mean1} - \text{mean0}) / s_p$, where s_p is the pooled standard deviation.

Result (Welch t-test): mean BMI 28.5661 (cardio=1) vs 26.5482 (cardio=0), $t=44.4318$, $p \approx 0$, Cohen’s $d = 0.3359$.

Hypothesis 2 (systolic BP difference by CVD status):

Welch two-sample t-test

Result (Welch t-test): mean ap_hi **133.8861** (cardio=1) vs **119.6029** (cardio=0), $t=123.7266$, $p \approx 0$, Cohen's $d = 0.9475$.

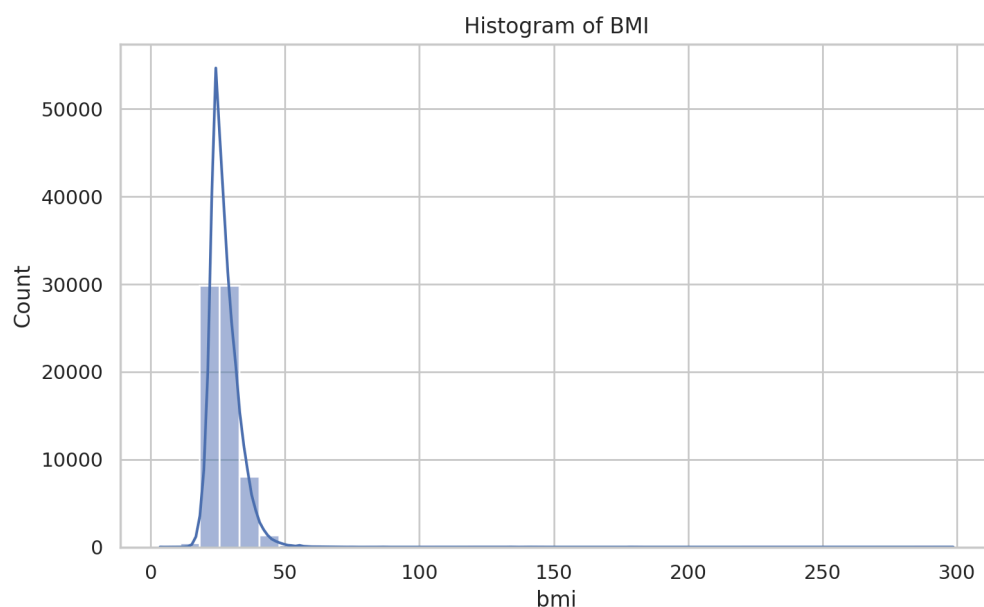
Hypothesis 3 (systolic BP differs by cholesterol category):

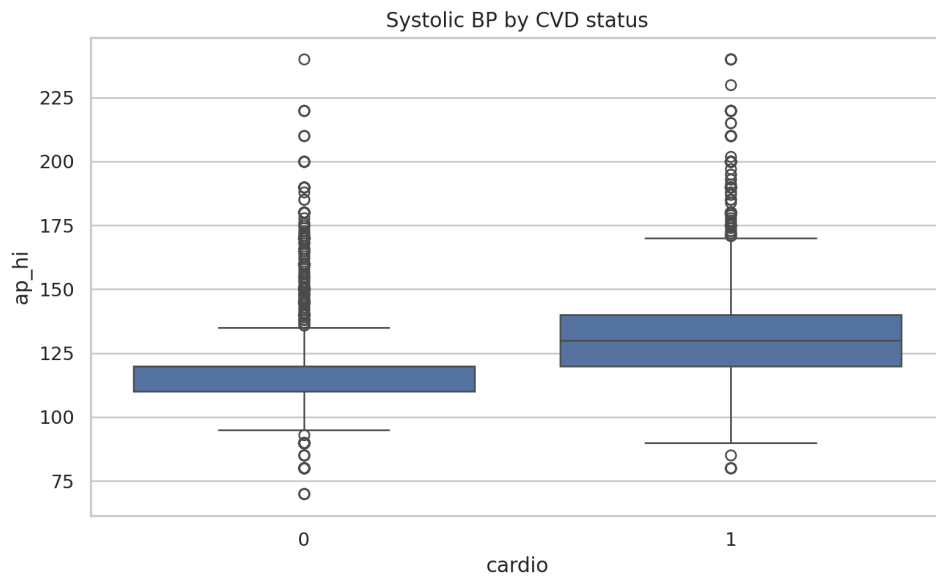
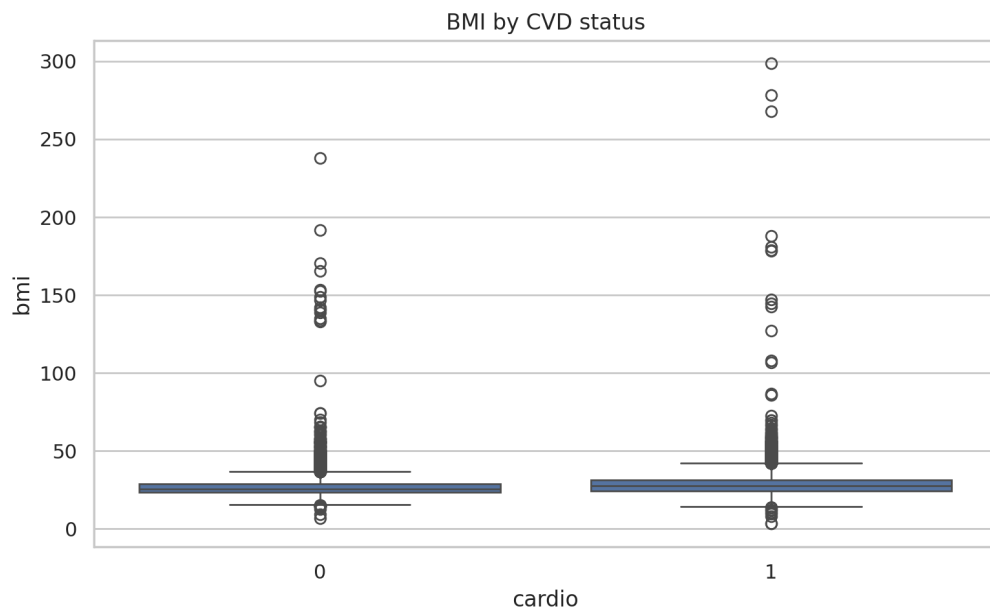
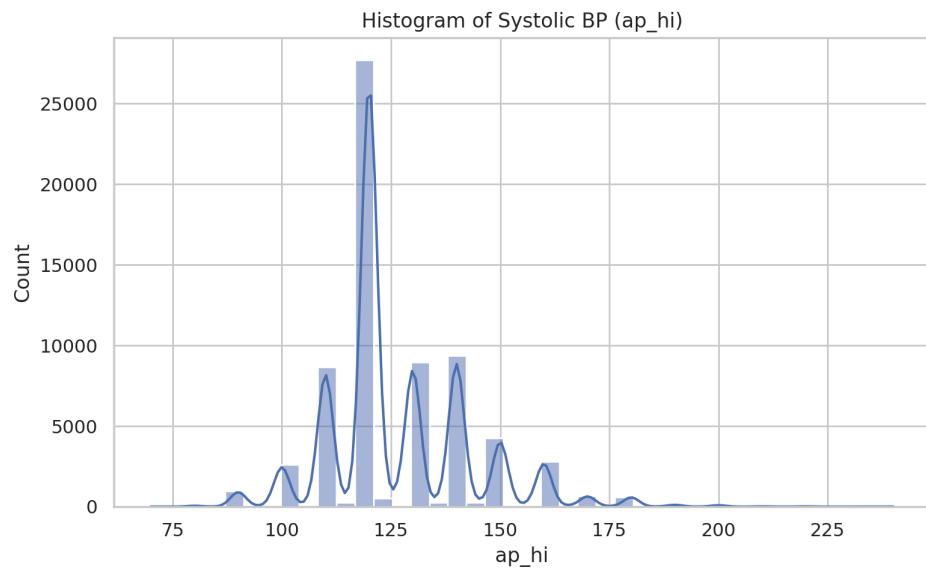
$H_0: \mu_1 = \mu_2 = \mu_3$ versus H_1 : at least one mean differs.

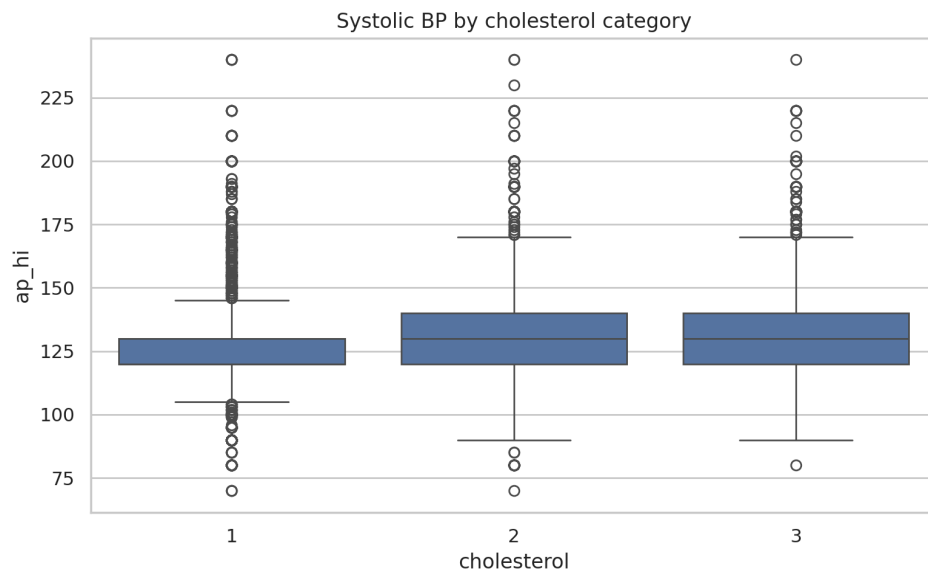
F statistic: $F = MS_{\text{between}} / MS_{\text{within}}$, and effect size $\eta^2 = SS_{\text{between}} / SS_{\text{total}}$.

Result (ANOVA): $F = 1393.6225$, $p \approx 0$, $\eta^2 = 0.0390$, with mean ap_hi increasing from **124.8153** (chol=1) to **130.8637** (chol=2) to **133.8409** (chol=3).

Interpretation (association strength, not causality): systolic BP shows a substantially larger standardized separation between CVD vs non-CVD than BMI ($d \approx 0.95$ vs 0.34), indicating stronger association with the outcome in this dataset.







Correlation analysis

Continuous–continuous associations (Pearson/Spearman/Kendall).

Pair	N	Pearson r (p)	Spearman ρ (p)	Kendall τ (p)
Age vs BMI	70,000	0.0855 (9.80e-114)	0.1076 (2.56e-179)	0.0721 (8.45e-180)
Age vs SBP (ap_hi)	68,668	0.2093 (≈0)	0.2227 (≈0)	0.1630 (≈0)
BMI vs SBP (ap_hi)	68,668	0.2333 (≈0)	0.2812 (≈0)	0.2076 (≈0)
SBP vs DBP	68,668	0.7346 (≈0)	0.7430 (≈0)	0.6689 (≈0)

Binary–continuous associations (point-biserial):

Cardio vs BMI:

$$r_{pb} = 0.1656, p \approx 0.$$

Cardio vs systolic BP:

$$r_{pb} = 0.4281, p \approx 0.$$

Active vs systolic BP:

$r_{pb} = -0.0011$, $p = 0.7679$ (no meaningful association in this sample for this definition).

Binary–binary associations (Phi):

Smoke vs cardio:

$\phi = -0.0155$, $p = 4.18e-05$ (small magnitude).

Active vs cardio:

$\phi = -0.0357$, $p = 3.99e-21$ (small magnitude).

Smoke vs alcohol:

$\phi = 0.3401$, $p \approx 0$ (moderate association between the behaviors).

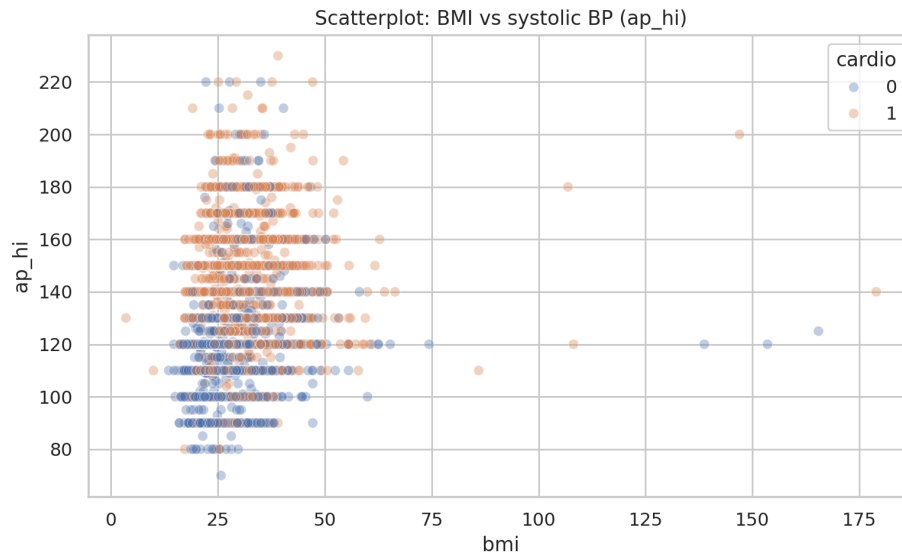
Joint “impact” analysis: hypertension vs obesity vs both

To support an interpretable clinical narrative, obesity was defined as BMI ≥ 30 and hypertension as systolic ≥ 140 or diastolic ≥ 90 (BP-plausible subset). The observed CVD prevalence by exposure pattern is: Neither **0.3205**, Obese-only **0.4490**, Hypertensive-only **0.7702**, Both **0.7958**.

Relative to the “Neither” group, association-strength measures are:

- Obese only vs Neither: RR = 1.4009 (95% CI 1.3629–1.4399), OR = 1.7275 (95% CI 1.6477–1.8112), risk difference = 0.1285.
- Hypertensive only vs Neither: RR = 2.4031 (95% CI 2.3616–2.4454), OR = 7.1064 (95% CI 6.7964–7.4306), risk difference = 0.4497.
- Both vs Neither: RR = 2.4828 (95% CI 2.4379–2.5285), OR = 8.2611 (95% CI 7.8149–8.7327), risk difference = 0.4753.

Comparing “Both” to “Hypertensive-only,” the absolute prevalence difference is **0.0256** ($0.7958 - 0.7702$), indicating that—within this dataset—adding obesity on top of hypertension produces a smaller incremental increase than the increase associated with hypertension alone relative to neither. On the odds-ratio interaction check, OR_{both} divided by $OR_{obese} \times OR_{hypertensive}$ is **0.6729**, suggesting overlap rather than strictly multiplicative association on the odds scale.



INTERPRETATION, REAL-WORLD APPLICATION, AND CONCLUSION

Interpretation

Across multiple association metrics, systolic blood pressure (SBP) shows stronger association with CVD status than BMI in this dataset. Specifically, the standardized separation between $cardio=1$ and $cardio=0$ is large for SBP (Cohen's $d = 0.9475$) and smaller for BMI (Cohen's $d = 0.3359$), and the point-biserial correlations show the same pattern ($r_{pb} = 0.4281$ for SBP vs $r_{pb} = 0.1656$ for BMI). Therefore, within this cross-sectional exam dataset, SBP appears to be the higher-yield marker of CVD status compared to BMI when “impact/toxicity” is defined strictly as magnitude of association.

Clinical meaning of the joint exposure analysis

When obesity is defined as BMI ≥ 30 and hypertension as SBP ≥ 140 or DBP ≥ 90 , the observed CVD prevalence rises from 0.3205 in the “Neither” group to 0.4490 in the “Obese-only” group and to 0.7702 in the “Hypertensive-only” group, with the highest prevalence in the “Both” group (0.7958). Relative to “Neither,” obesity alone corresponds to RR = 1.4009 and OR = 1.7275, whereas hypertension alone corresponds to RR = 2.4031 and OR = 7.1064, again indicating substantially stronger association for hypertension than obesity under these definitions. The incremental difference between “Both” and “Hypertensive-only” is 0.0256 in absolute prevalence (0.7958 - 0.7702), suggesting that—once hypertension is present—obesity adds comparatively smaller additional association with CVD status in this dataset (though both exposures still represent a high-risk profile).

Implications for screening and prevention

Prioritization for follow-up: In a clinical workflow, patients meeting the hypertension definition should be prioritized for near-term evaluation and monitoring because the association with CVD status is markedly larger (RR 2.4031 and OR 7.1064 vs “Neither”) than the association for obesity alone (RR 1.4009 and OR 1.7275). This does not imply causality, but it supports hypertension as a strong flag for risk stratification within similar exam-based datasets.

Dual-risk messaging: The “Both” group exhibits the highest observed CVD prevalence (0.7958), so patients who are both hypertensive and obese should be treated as a high-priority prevention group for combined risk-factor management (blood pressure control plus weight management), even if the incremental prevalence increase above hypertension alone is modest.

Integrated cardiometabolic context: The ANOVA results show that higher cholesterol category is associated with higher mean SBP (from 124.8153 in cholesterol=1 to 133.8409 in cholesterol=3), supporting the interpretation that risk factors cluster and that multi-factor clinical counseling (BP control alongside lipid management) is appropriate.

Consistency with the broader evidence base

The pattern observed here is consistent with global research emphasizing hypertension and obesity among major modifiable risk factors associated with myocardial infarction and cardiovascular outcomes (e.g., INTERHEART), while also highlighting that risk factors commonly co-occur rather than act in isolation. The correlations among age, BMI, and blood pressure (e.g., BMI vs SBP correlations positive across Pearson/Spearman/Kendall) further support this clustering view and justify analyzing combined exposure groups rather than only single-variable comparisons.

Limitations

This analysis is cross-sectional, so associations cannot be interpreted as causal effects, and unmeasured confounding (e.g., medications, prior diagnosis, diet, socioeconomic factors) may influence observed relationships. In addition, binary cutoffs (BMI ≥ 30 ; SBP/DBP thresholds) simplify continuous biology and may change effect estimates if alternative thresholds are used. Despite these constraints, the large sample size and consistent ranking across effect sizes (SBP > BMI) provide a clear and practically interpretable conclusion about association strength in this dataset.

Conclusion

Within this dataset, systolic blood pressure demonstrates stronger association with CVD status than BMI, and the coexistence of hypertension and obesity corresponds to the highest observed CVD prevalence. For real-world screening and prevention prioritization based on association strength alone, hypertension is the most informative single marker among the two, while combined risk remains clinically meaningful as a high-prevalence group.