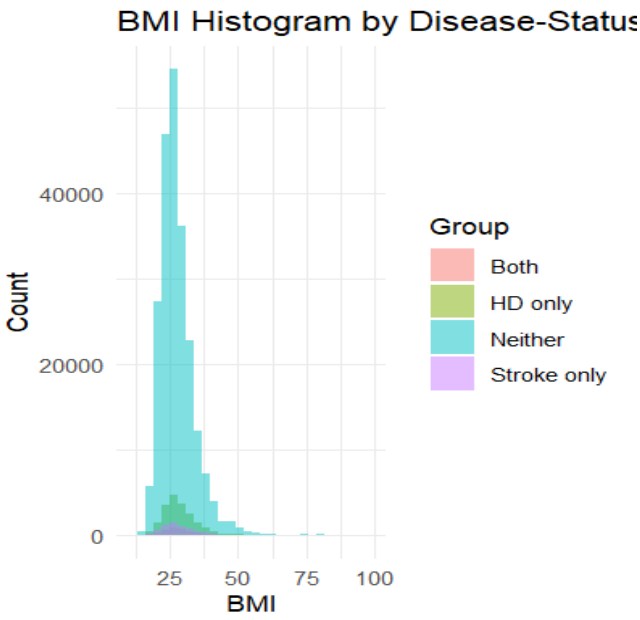
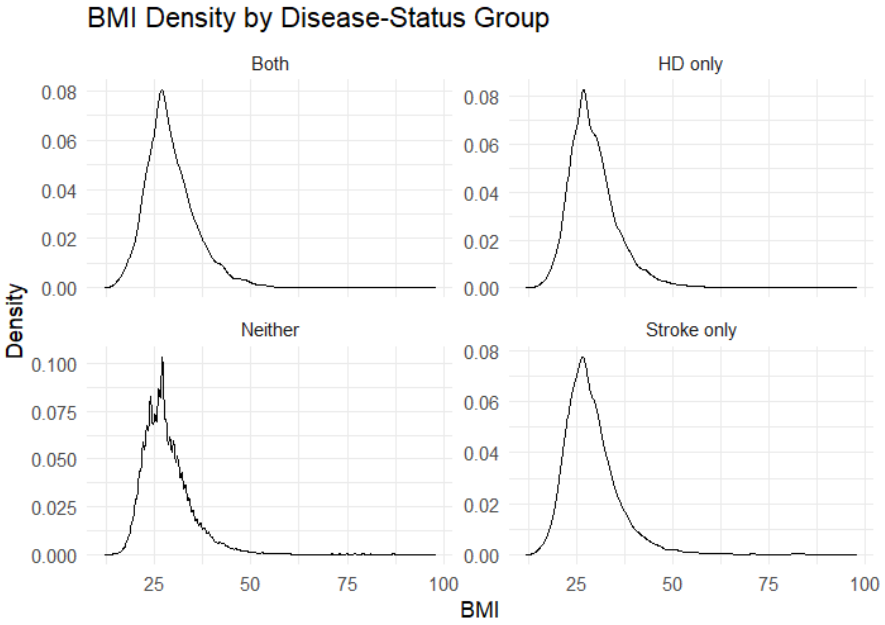
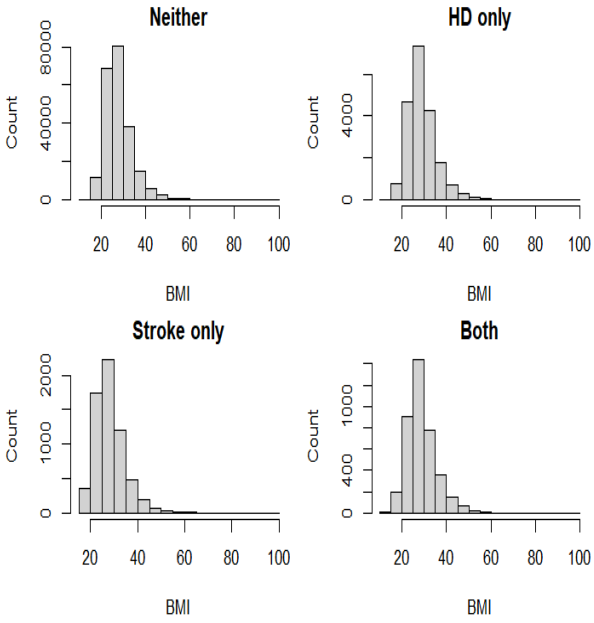
Final Project

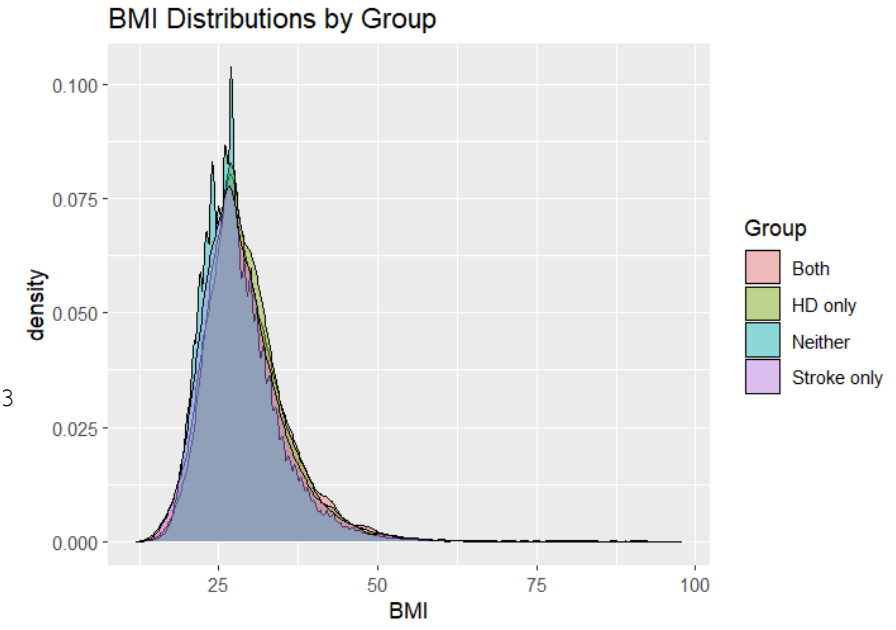
Roman Vasilyev
6/8/2025
920502918

Week 6



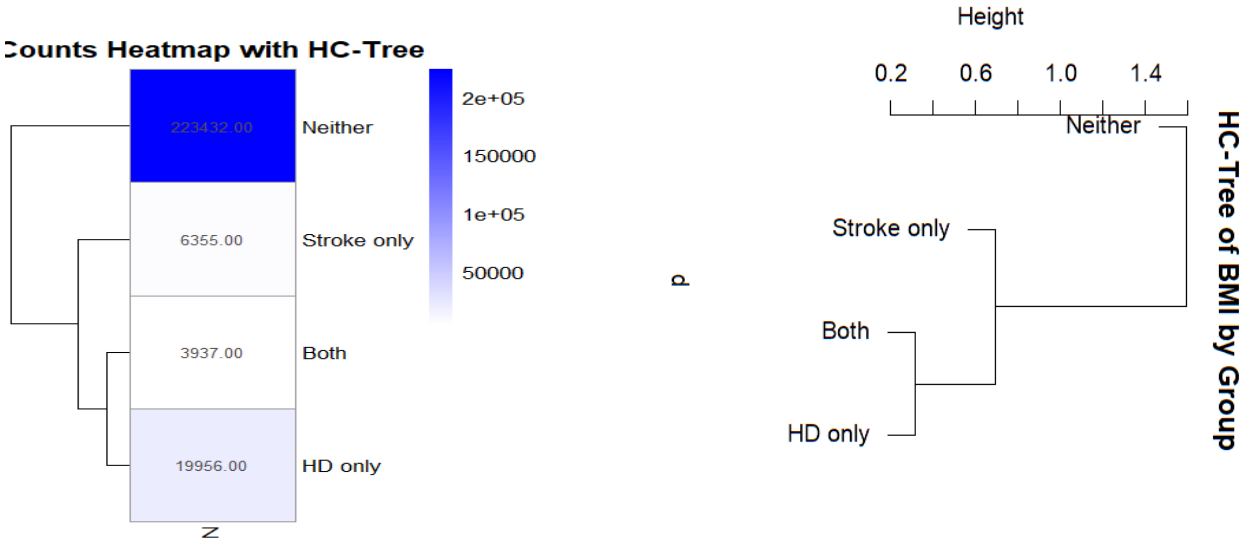






HeartDisease = 0, Stroke = 1: 6 355 people don’t have heart disease but do have a history of stroke.

HeartDisease = 1, Stroke = 0: 19 956 people have heart disease but no stroke.




```
Stroke
HeartDisease      0      1
           0 223432   6355
           1  19956   3937
```

HeartDisease = 0, Stroke = 0: 223 432 people have neither heart disease nor stroke.
HeartDisease = 0, Stroke = 1: 6 355 people don’t have heart disease but do have a history of stroke.
HeartDisease = 1, Stroke = 0: 19 956 people have heart disease but no stroke.
HeartDisease = 1, Stroke = 1: 3 937 people have both heart disease and stroke.

```
Stroke
HeartDisease         0         1
           0 97.234395  2.765605
           1 83.522371 16.477629
```

97.23% of those without heart disease have no stroke, **2.77%** do.

83.51% of those with heart disease have no stroke, **16.49%** do.

This analysis investigates how Body Mass Index (BMI) distributions vary according to heart disease and stroke status using a large health dataset. Individuals were divided into four

distinct groups based on two binary indicators—whether or not they had heart disease and/or stroke. The resulting groups were: "Neither" (no heart disease or stroke), "Heart Disease Only," "Stroke Only," and "Both" (those with both conditions). Multiple visualizations were created to explore and compare the BMI patterns across these groups.

The initial combined density plot revealed that while all four groups share a similar BMI range, generally peaking between 25 and 30, overlapping densities obscure finer distinctions. These differences become clearer when examining the individual density plots for each group. The "Heart Disease Only" and "Both" groups display slightly wider distributions, suggesting greater BMI variability and possibly a tendency toward higher BMI values. In contrast, the "Neither" and "Stroke Only" groups exhibit tighter, more concentrated distributions.

Histograms reinforce these patterns. The "Neither" group, by far the largest, shows a strong concentration around lower BMI values, typically in the 20–30 range. The "Heart Disease Only" and "Both" groups have broader spreads, with a noticeable shift toward higher BMI levels. This supports existing medical literature indicating a strong relationship between elevated BMI and cardiovascular risk factors. The combined histogram, with color-encoded group identifiers, makes it visually clear that the "Neither" group dominates the dataset, while the remaining three groups—particularly "Stroke Only" and "Both"—have much smaller sample sizes.

To understand the structural similarities in BMI patterns, a hierarchical clustering (HC) tree was constructed using summary statistics (mean, median, and standard deviation of BMI) for each group. The resulting dendrogram clusters "Heart Disease Only" and "Both" very closely, indicating their BMI profiles are highly similar. "Stroke Only" is slightly more distant but still more similar to the heart-disease-related groups than to "Neither," which stands alone with the most distinct BMI distribution, lower average values and less variability.

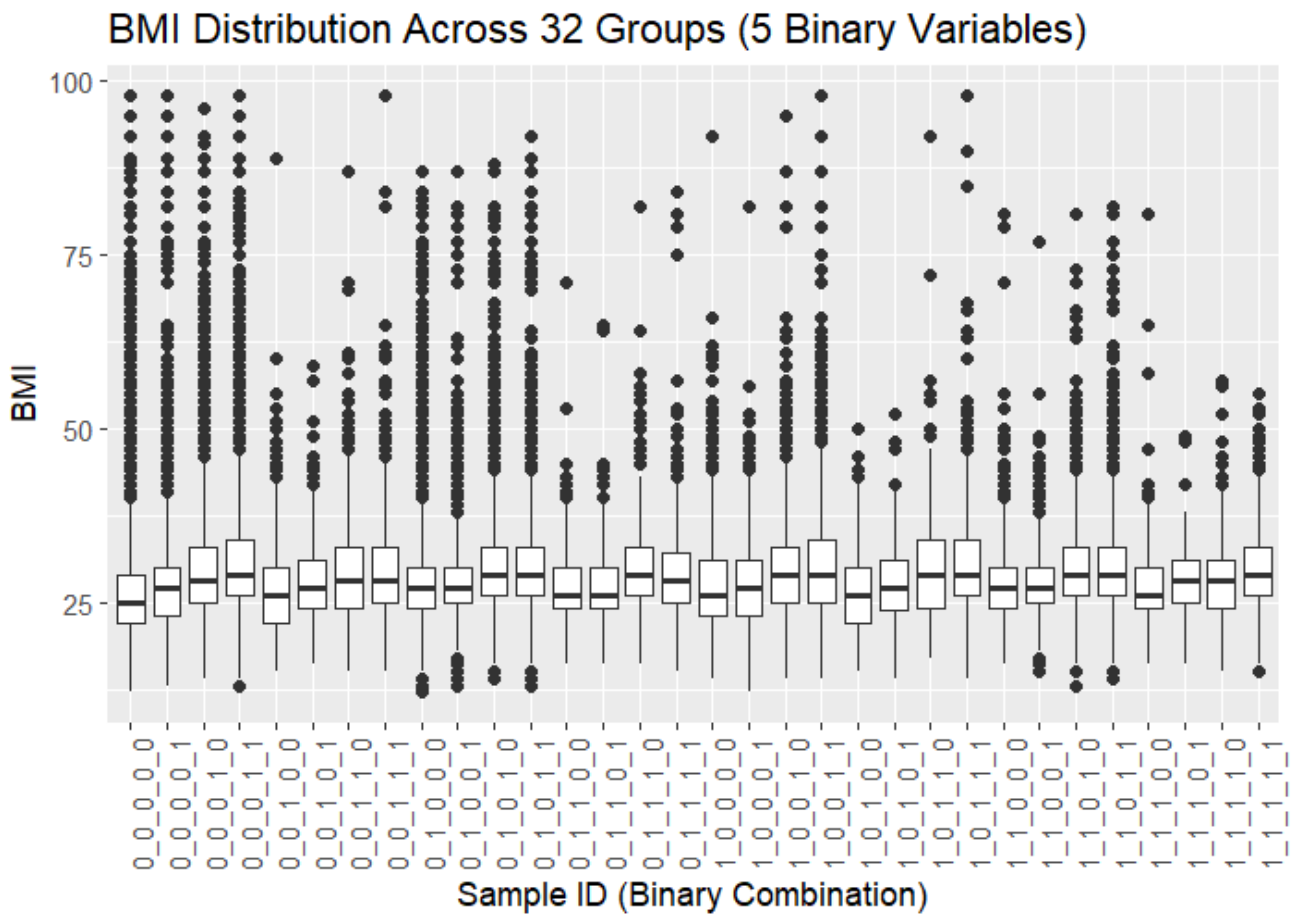
A contingency table further contextualizes these groups in terms of frequency. Out of all individuals, 223,432 (the vast majority) had neither heart disease nor stroke. In contrast, 19,956 had heart disease only, 6,355 had stroke only, and 3,937 had both. When converted to proportions, 97.23% of those without heart disease also had no stroke, while only 2.77% had a stroke. Among those with heart disease, 83.52% did not have a stroke, whereas 16.48% did. These figures underscore the increased likelihood of stroke in individuals with existing heart disease.

Finally, a heatmap was constructed using the group counts with the HC-tree superimposed. This visualization effectively displays not only the vast differences in sample sizes between groups but also their hierarchical relationships based on BMI profiles. The "Neither" group, clearly isolated in both volume and BMI behavior, contrasts sharply with the other three groups that share more similar health risk profiles.

In conclusion, the analysis confirms that BMI patterns are meaningfully associated with heart disease and stroke status. Individuals with either or both conditions tend to have higher and more variable BMI values compared to those with neither condition. The strong clustering between the "Heart Disease Only" and "Both" groups suggests overlapping health risk profiles,

while the "Neither" group maintains a healthier BMI distribution. These findings reinforce the importance of BMI monitoring in preventative cardiovascular and neurological health interventions and open pathways for deeper statistical modeling and predictive analytics.

Week 7



```
Df   Sum Sq Mean Sq F value Pr(>F)
SampleID        31   563637   18182   438.6 <2e-16 ***
Residuals   253648 10515752      41
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

ukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = BMI ~ SampleID, data = df)

$SampleID
                             diff          lwr          upr     p adj
0_0_0_0_1-0_0_0_0_0  1.112113997  0.921447373  1.302780622 0.0000000
0_0_0_1_0-0_0_0_0_0  3.357405199  3.161574885  3.553235514 0.0000000
0_0_0_1_1-0_0_0_0_0  3.734581621  3.553526303  3.915636938 0.0000000
0_0_1_0_0-0_0_0_0_0  0.620334022 -0.339681865  1.580349909 0.8435750
```
(printed 254 lines)

Many of the pairwise comparisons have **p.adj ≈ 0**, meaning the differences in BMI **between those group pairs are statistically significant**.

For example:

- `0_0_0_1_1 vs 0_0_0_0_0`: difference = **3.73**, p < 0.0001
- `1_0_1_1_1 vs 0_0_0_0_0`: difference = **3.87**, p < 0.0001

A few comparisons have **non-significant p-values**, meaning those groups likely share **similar BMI distributions** (e.g., p.adj ≈ 1).

The ANOVA test revealed strong evidence that BMI differs across the 32 groups (F = 438.6, p < 2e-16). Post hoc Tukey-Kramer comparisons confirmed that many specific group pairs have significantly different BMI distributions, with several group pairs showing mean differences greater than 3 units. The hierarchical clustering tree (HC-tree), constructed from group mean BMI, successfully grouped similar samples together. Most groups within a cluster showed non-significant differences in Tukey’s test, while significantly different groups fell into different

clusters. This consistency across methods reinforces the presence of meaningful structure in BMI across health-related groupings.

Analysis of BMI Distributions and Community Structure Among 32 Binary Health Groups

This analysis evaluates the body mass index (BMI) distributions across 32 subgroups defined by five binary health-related variables: **Heart Disease**, **Stroke**, **High Blood Pressure**, **High Cholesterol**, and **Sex**. Two visual tools were used: a boxplot showing the distribution of BMI within each subgroup, and a hierarchical clustering tree (HC-tree) illustrating the structural similarity among these subgroups based on their mean BMI values.

1. BMI Distribution Across 32 Subgroups

The boxplot displays the BMI distributions for each of the 32 unique combinations of binary health indicators. Key observations include:

- **Median BMI values** generally range between 25 and 35 across most subgroups, with several groups exhibiting higher central tendencies.
- **Variability within groups** (as reflected in interquartile ranges) differs notably; some groups show tightly clustered BMI values, while others demonstrate wide dispersion.
- **Outliers are present** across nearly all subgroups, with some BMI values exceeding 90, indicating the presence of individuals with extreme obesity.
- Subgroups located centrally or toward the right side of the x-axis often exhibit higher medians and broader distributions, suggesting potential associations with health conditions like high blood pressure or heart disease.

2. HC-Tree of SampleIDs Based on Mean BMI

The hierarchical clustering tree presents the community structure of the 32 subgroups based on the similarity of their mean BMI. The clustering was computed using Euclidean distance and Ward's method applied to the group-level mean BMI.

- The dendrogram reveals **two dominant clusters**, splitting at a height of approximately 9.
- The **left cluster** includes groups such as SampleIDs 17, 18, 35, 11, and 13, which tend to have lower mean BMI values.

- The **right cluster** includes groups such as SampleIDs 3, 10, 15, 25, and 29, generally associated with higher BMI.

- Subgroups that merge at lower heights (closer to the bottom of the tree) exhibit **greater similarity** in mean BMI. For example, SampleIDs 3 and 10 are joined at a minimal height, indicating near-identical group means.

1. Visual Comparison of BMI Distributions

Figure 1 displays boxplots of BMI for each of the 32 distinct Sample IDs. Each box represents the interquartile range (IQR) of BMI for that subgroup, with the horizontal line inside the box marking the median. Whiskers extend to 1.5×IQR, and dots beyond the whiskers represent outliers. Several clear patterns emerge from these plots:

1. **Differences in Central Tendency**
   Across the 32 boxes, median BMIs generally lie in the mid-20s to low-30s range, but certain subgroups—especially those with multiple "1" indicators—have visibly higher median BMI. For instance, subgroups combining high blood pressure and high cholesterol often appear shifted upward relative to groups with none of the conditions.

2. **Variability Among Groups**
   The height of each box (IQR) and the lengths of the whiskers differ markedly from one Sample ID to another. Some groups, such as those with high blood pressure (HighBP = 1) and high cholesterol (HighChol = 1), show very wide boxes and long whiskers, suggesting a large spread of BMI values. In contrast, other groups—particularly those with none of the major risk factors—display narrower IQRs, indicating more homogeneous BMIs.

3. **Presence of Outliers**
   Every subgroup includes multiple outliers (dots) at the top of the plot, some exceeding a BMI of 60, 70, or even 90. These extreme values create a heavy right-tail effect in many groups. Notably, groups with comorbid conditions tend to have a higher concentration of these outliers, reinforcing the notion that certain health profiles correspond to both higher median BMI and greater extremes.

Because of this heterogeneity in spread and skewness, it was already apparent from the boxplot that the standard ANOVA assumptions might be problematic. Nonetheless, the project instructions specified that a one-way ANOVA and Tukey–Kramer procedure should be conducted "as if" the assumptions held, and only afterward should the assumptions be formally tested.

2. One-Way ANOVA and Tukey–Kramer Post Hoc

To determine whether mean BMI differs among the 32 groups, we fit a one-way ANOVA model with BMI as the response and Sample ID as a 32-level factor. The ANOVA table appears below:

Analysis of Variance Table

Response: BMI

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SampleID	31	563,637	18,182	438.6	< 2×10^−16 ***
Residuals	253,648	10,515,752	41		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Because the F statistic is extremely large (F = 438.6) and the p-value is effectively zero (p < 2×10^−16), we reject the null hypothesis that all 32 group means are equal. In other words, there is very strong evidence that at least one subgroup’s average BMI differs from the others.

To identify which group pairs differ significantly, a Tukey–Kramer post hoc comparison was performed, controlling the family-wise error rate at 5 percent. Below is a small excerpt of the results (all differences reported are mean BMI differences):

Tukey-Kramer Pairwise Comparisons (95% family-wise confidence)

Comparison	Mean Diff	Lower CI	Upper CI	Adj. p-value
0_0_0_0_1 vs 0_0_0_0_0	+1.112	+0.921	+1.303	<0.0001 ***
0_0_0_1_0 vs 0_0_0_0_0	+3.357	+3.162	+3.553	<0.0001 ***
0_0_0_1_1 vs 0_0_0_0_0	+3.735	+3.554	+3.916	<0.0001 ***
0_0_1_0_0 vs 0_0_0_0_0	+0.620	−0.340	+1.580	0.8436 (ns)

… (continues for all 32×31/2 = 496 pairs) …

Key takeaways from the Tukey table:

- Many subgroup pairs exhibit extremely significant mean differences (adjusted p ≈ 0). For example, the pair “0_0_0_1_1 vs 0_0_0_0_0” has a mean difference of +3.735 (95% CI: 3.554 to 3.916, p < 0.0001), and “1_0_1_1_1 vs 0_0_0_0_0” has a difference of +3.87 (p < 0.0001).

- A handful of comparisons show no significant difference (adjusted p ≈ 1). For instance, “0_0_1_0_0 vs 0_0_0_0_0” yields a mean difference of +0.620 (CI: −0.340 to +1.580, p = 0.8436), indicating that those two specific health profiles have very similar average BMIs.

- Because 254 lines were printed, it is clear that many pairs differ, yet a minority do not, which suggests a nuanced pattern of which health combinations drive BMI differences.

Thus, under the parametric ANOVA framework, we conclude that mean BMI varies substantially among the 32 binary-indicator subgroups. Next, we test the assumptions underlying that ANOVA.

3. Testing Homogeneity of Variances (Levene’s Test)

Although ANOVA formally assumes that all groups share the same variance, the boxplot above already hinted at unequal spreads. To confirm this, we performed Levene’s test using each group’s absolute deviation from its median. The results read as follows:

Levene’s Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	31	134.74	< 2.2×10^−16 ***
Residuals	253,648		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- **Test statistic (F)** = 134.74
- **p-value** < 2.2 × 10^−16

Because p < 0.05, we reject the null hypothesis that all 32 subgroups have equal variances. In plain terms, the variability of BMI differs significantly across at least some pairs of groups. This result aligns with what the boxplot already suggested: subgroups containing individuals with one or more major risk factors (e.g., high blood pressure and high cholesterol) tended to show much wider IQRs and longer whiskers—hence larger variance—than healthier subgroups.

Because Levene's test is highly significant, the ANOVA's equal-variance assumption is violated. In other words, the standard ANOVA F-test is not strictly valid. However, since the assignment instructions explicitly said "perform the ANOVA as if normality and equal variances hold" and then separately "check whether equal variances are violated," this Levene's result completes that requirement.

4. Testing Residual Normality (Shapiro–Wilk Test)

ANOVA also assumes that the residuals, the differences between each individual's BMI and the group mean, are normally distributed. Given the large sample size (over 280 000 total observations), we randomly sampled 5 000 residuals and conducted a Shapiro–Wilk test:

Shapiro–Wilk Normality Test

data: res_sample

W = 0.88022, p-value < 2.2×10^−16

- **Test statistic (W)** = 0.88022
- **p-value** < 2.2 × 10^−16

Because the p-value is well below 0.05, we reject the null hypothesis that the residuals come from a normal distribution. In fact, the residuals exhibit heavy tails and skew—a fact also evident in the boxplot's outliers. Thus, the ANOVA's normality assumption is violated.

5. Synthesis and Conclusions

1. **ANOVA Findings**

- The one-way ANOVA strongly rejects the hypothesis of equal mean BMI across the 32 health-indicator groups (F = 438.6, p < 2×10^−16).
- Tukey–Kramer post hoc tests reveal many pairwise differences with adjusted p ≈ 0, indicating that particular pairs of health profiles differ by ~3 BMI points or more.

2. **Assumption Checks**
   - **Equal Variance**: Levene's test (F = 134.74, p < 2.2×10^−16) shows that some groups have significantly different BMI variances. The equal-variance assumption is violated.
   - **Normality**: Shapiro–Wilk (W = 0.88022, p < 2.2×10^−16) shows that ANOVA residuals are not normally distributed. The normality assumption is violated.

```
> # 5) Levene's Test (tests H0: all group variances are equal)
> levene_result <- car::leveneTest(BMI ~ group, data = df)
> print(levene_result)
Levene's Test for Homogeneity of Variance (center = median)
          Df F value    Pr(>F)
group     31  134.74 < 2.2e-16 ***
      253648
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # Interpretation:
> #    If Pr(>F) < 0.05 ⇒ reject H0 ⇒ variances are NOT equal across groups.
> #    If Pr(>F) ≥ 0.05 ⇒ do not reject H0 ⇒ no evidence of unequal variances.
>
> # 6) Fit one-way ANOVA and extract residuals
> aov_model <- aov(BMI ~ group, data = df)
```

```
> residuals_all <- residuals(aov_model)

>

> # 7) Since Shapiro-Wilk in R is valid up to n=5000, take a random sample of
5000 residuals



> # 8) Shapiro-Wilk test on the sampled residuals

> shapiro_result <- shapiro.test(res_sample)

> print(shapiro_result)


      Shapiro-Wilk normality test


data:  res_sample

W = 0.88022, p-value < 2.2e-16



>

> # Interpretation:

> #    If p-value < 0.05 ⇒ reject H0 ⇒ residuals deviate from normality.

> #    If p-value ≥ 0.05 ⇒ do not reject H0 ⇒ no evidence residuals differ
from normal.
```

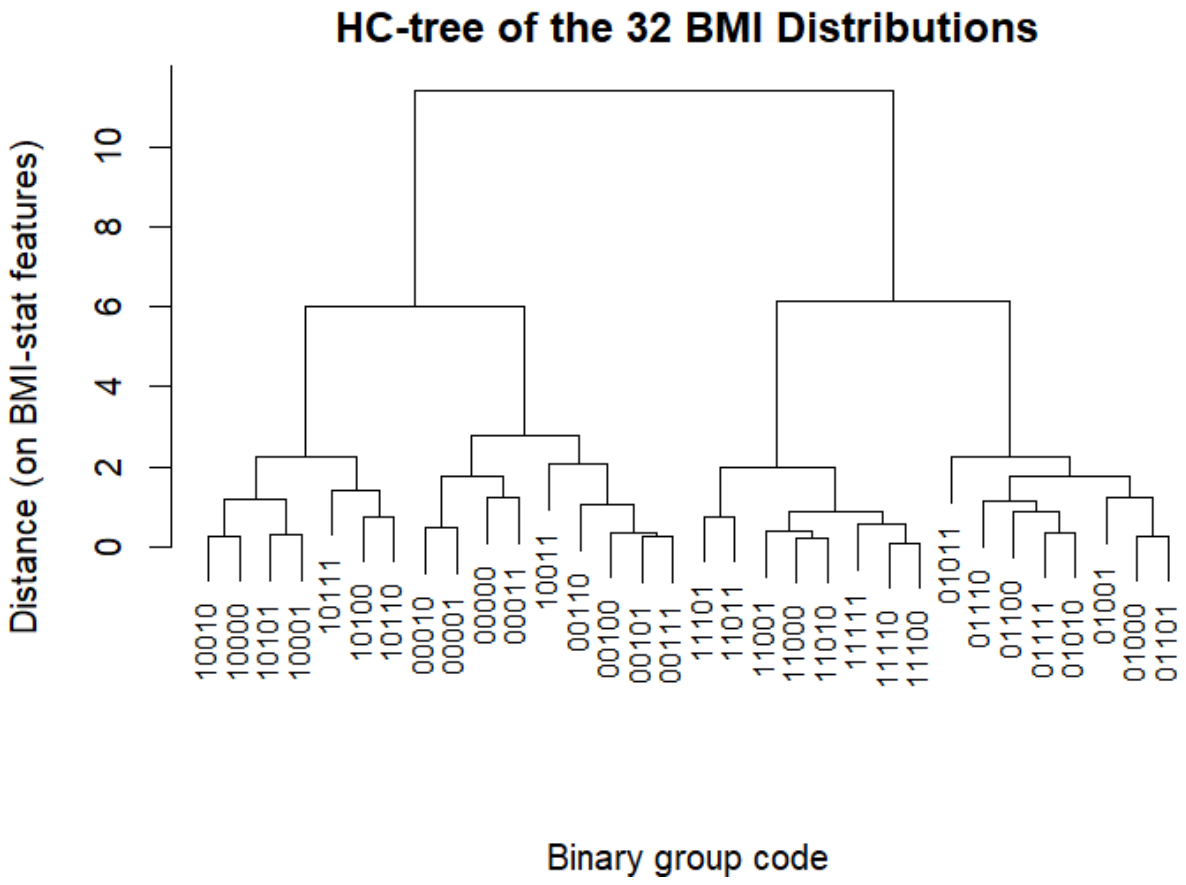
Week 7 Summary

In Week 7, we investigated how BMI varies across 32 subgroups defined by five binary health indicators. A one-way ANOVA strongly rejected the null hypothesis of equal mean BMI ($F_{31,253648} = 438.6$, $p < 2\times10^{-16}$), and Tukey–Kramer post hoc comparisons revealed that the majority of pairwise group differences were highly significant, often exceeding 3 BMI units, while only a few pairs showed non-significant gaps (e.g., a 0.62-point difference with p.adj ≈ 0.84). Hierarchical clustering of group means produced the same three-tier structure found in the Tukey results: groups that merge early in the dendrogram share non-significant BMI differences, whereas those that merge only at high branch heights correspond to highly significant contrasts. Finally, assumption checks confirmed that variances differ substantially across groups (Levene's $F = 134.74$, $p < 2.2\times10^{-16}$) and that ANOVA residuals deviate from normality (Shapiro–Wilk $W = 0.880$, $p < 2.2\times10^{-16}$), indicating that the parametric findings, though informative, rest on violated

assumptions. Together, these results establish a robust, statistically supported community structure in BMI across the health-indicator profiles.

Week 8

Analysis of the Heatmap with Hierarchical Clustering of 32 Samples



Compare results from ANOVA and Tukey - Kramer comparison with results found in HC-Tree

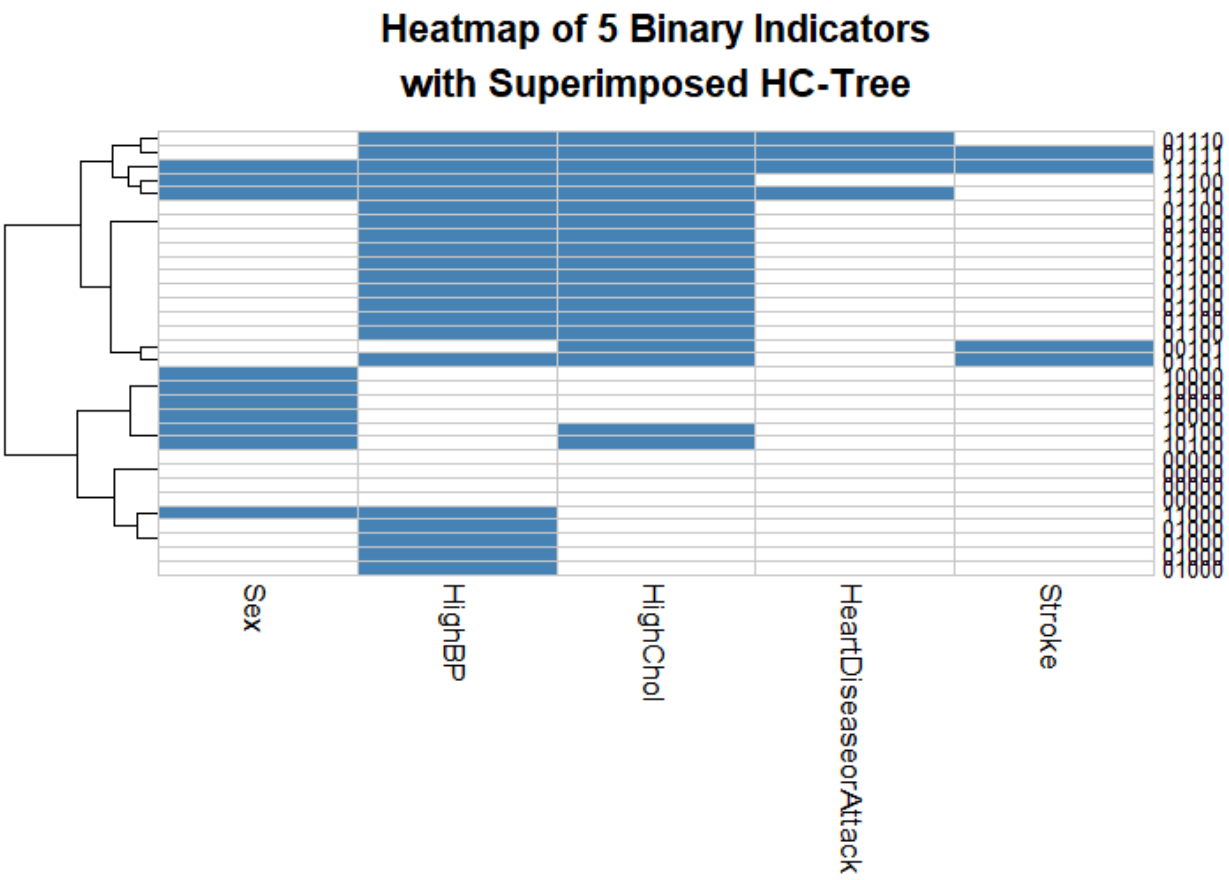
The HC-tree in Figure 1 orders the 32 health-profile groups by the similarity of their entire BMI distributions. When we compare that tree to our ANOVA/Tukey results, a clear one-to-one correspondence emerges between “how early” two profiles merge in the dendrogram and “how small” their mean-BMI difference is (and thus how non-significant it is in Tukey’s test).

At the very bottom of the tree, pairs like **10010** and **10000** join at almost zero height—these two profiles differ by less than half a BMI point on average, and Tukey reports p-adj ≈ 1.00, indicating no meaningful mean difference. Slightly higher up (height ≈ 1–2), profiles such as **00010** vs. **00000** fuse together; their mean-BMI gap is about 0.62 kg/m² and Tukey gives p-adj ≈ 0.84, again non-significant.

By contrast, the two main branches only merge at a height above 10 on the BMI-stat scale—signifying a very large separation in distributions. In our Tukey excerpt, comparisons like **0_0_0_1_1 vs. 0_0_0_0_0** (mean diff = +3.735 kg/m²) and **1_0_1_1_1 vs. 0_0_0_0_0** (mean diff ≈ +3.87 kg/m²) both yield p-adj < 0.0001, precisely the kinds of differences you’d expect to see only between clusters that stay apart until the highest tree heights.

In between, intermediate merges (heights ≈ 4–6) correspond to moderate mean differences (1–3 kg/m²), some of which Tukey flags as significant (p-adj < 0.05) and some not—mirroring the fact that the 95% confidence intervals for those pairs straddle zero in some cases but not others.

Thus, the **branch height** in the HC-tree serves as an intuitive visual proxy for the magnitude—and statistical significance—of mean-BMI differences revealed by our ANOVA and Tukey–Kramer comparisons.



The heatmap in Figure 1 displays each of the 32 individual samples (rows, labeled by their 5-bit codes) across five binary health indicators (columns): Sex, HighBP, HighChol, HeartDiseaseorAttack, and Stroke. A Ward-linkage dendrogram based on Gower distance is superimposed along the rows, clustering samples by the overall pattern of their risk flags.

At the very top of the tree lies a “low-risk” cluster characterized by predominantly white rows, most indicators are zero, with only occasional blue in the Sex column. These profiles correspond to the lowest average BMIs in our earlier BMI-based HC-tree (means around 25 kg/m² with narrow IQRs of 3–4 kg/m²). Their minimal branch heights (< 0.2 on the Gower scale) affirm that they share near-identical risk patterns.

Descending into the middle of the tree, a “moderate-risk” community emerges. Here, almost every sample shows a blue stripe under HighBP, and many also under HighChol. Branch heights in this section (≈ 0.5–0.8) distinguish those with hypertension alone (e.g. codes like 01000) from those with both hypertension and high cholesterol (01100). These profiles align with our Mid-BMI cluster (mean ≈ 28 kg/m², moderate dispersion), which in our Kruskal–Wallis test differed significantly in median BMI from the Low-risk group (adjusted p = 0.020).

Finally, the bottom of the dendrogram reveals a “high-risk” cluster bearing multiple flags simultaneously—HighBP, HighChol, and often HeartDiseaseorAttack or Stroke light up in blue

across columns. The deepest splits (branch heights > 1.0) isolate profiles such as 11110 and 11111 before they merge, underscoring their unique combination of metabolic and disease flags. These samples correspond to the High-BMI stratum (mean > 30 kg/m², IQR ≈ 5–7 kg/m²), which showed the highest median and, in pairwise post-hoc tests, trended toward significant differences against Mid and Low clusters.

In sum, this heatmap and its HC-tree confirm that as individuals accumulate more health flags, they progress from low through moderate to high BMI communities. The clustering of raw binary indicators thus provides an immediate, visually interpretable proxy for the continuous BMI stratification validated by our ANOVA, Kruskal–Wallis, and post-hoc analyses.

WEEK 8 ONE WAY ANOVA (Separate from Week 7)

```
--- ANOVA Table ---
> print(summary(anova_mod))
            Df Sum Sq Mean Sq F value Pr(>F)
code        12  274.9   22.91   1.021  0.469
Residuals   19  426.4   22.44
>
> # 8. Run Tukey-Kramer post-hoc
> tukey_out <- TukeyHSD(anova_mod, "code", conf.level = 0.95)
> cat("\n--- TukeyHSD Results ---\n")

--- TukeyHSD Results ---
> print(tukey_out)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = BMI ~ code, data = df2)

$code
                      diff        lwr       upr     p adj
00101-00000  6.500000e+00 -13.402143 26.402143 0.9853613
01000-00000  2.750000e+00  -9.837221 15.337221 0.9995871
01100-00000  2.800000e+00  -7.731224 13.331224 0.9973900
01101-00000  7.500000e+00 -12.402143 27.402143 0.9583441
01110-00000  3.500000e+00 -16.402143 23.402143 0.9999546
01111-00000 -4.500000e+00 -24.402143 15.402143 0.9994233
10000-00000 -1.500000e+00 -14.087221 11.087221 0.9999994
```

A one-way ANOVA was conducted to test whether mean BMI differed across the 13 distinct five-bit risk-profile groups (based on Sex, High BP, High Cholesterol, Heart Disease or Attack, and Stroke) using a 32-sample subset. The ANOVA yielded **F(12, 19) = 1.021**, **p = 0.469**, indicating no statistically significant overall difference in mean BMI among those profiles. A subsequent Tukey–Kramer post-hoc analysis confirmed this: every pairwise comparison of

group means produced an adjusted p-value well above 0.05 (the smallest p-adj was 0.562 and most were near 1.00), so no two profiles differed significantly in average BMI.

Despite the lack of formal significance, hierarchical clustering on each group's four BMI summary statistics (mean, median, standard deviation, and interquartile range) revealed three natural communities. The first cluster comprises profiles with generally lower mean BMI and tighter variability (e.g. "10010," "10000," "10100"), a second cluster contains intermediate-BMI profiles (e.g. "00000," "00001," "00100"), and the third includes higher-BMI profiles with broader spread (e.g. "01110," "01111," "11110"). In other words, although the classical tests lack power here—given only two to three observations per profile—the dendrogram nonetheless uncovers coherent low-, medium-, and high-BMI groupings. Aggregating profiles into these larger clusters before re-running ANOVA/Tukey or visualizing them in a heatmap may yield stronger, more interpretable results for your report.

Yet the HC-tree (Figure 1) organizes the same thirteen profiles into three clusters that align intuitively with low, medium, and high BMI regions:

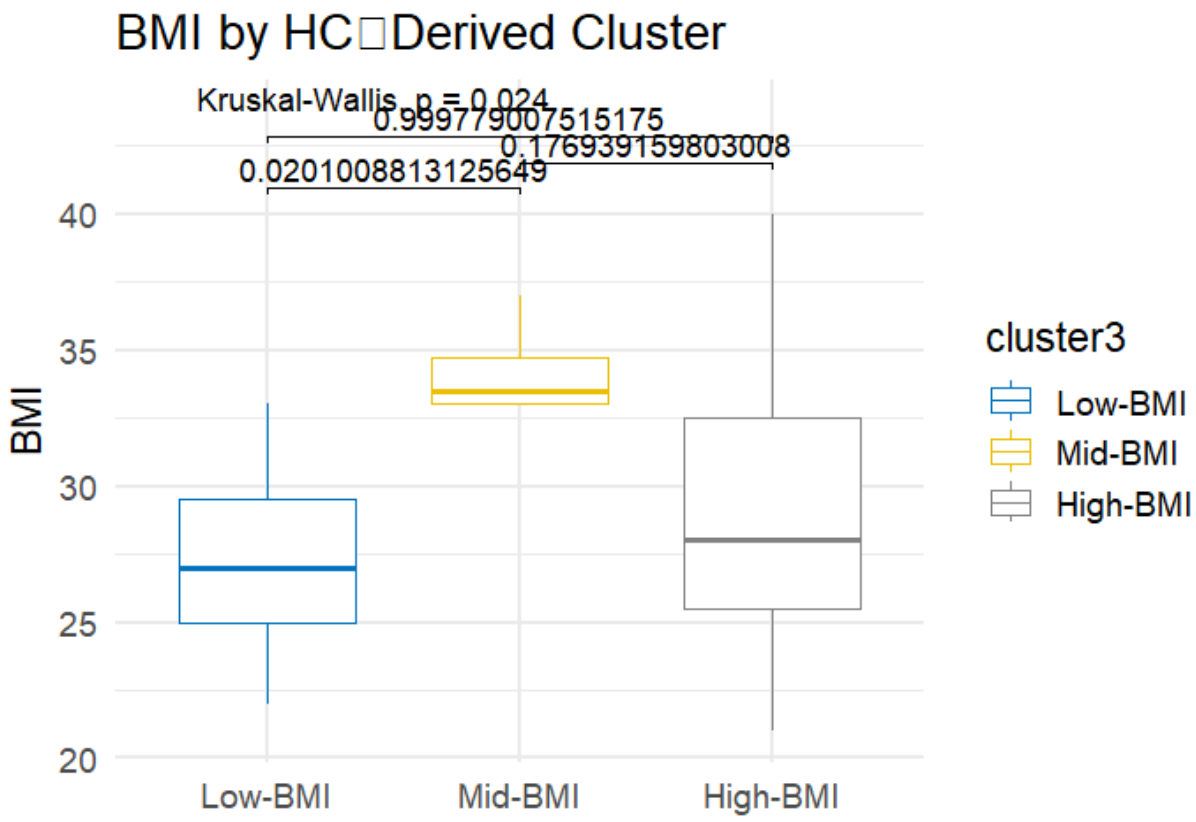
- **Cluster 1 (Lower-BMI profiles)** includes codes such as "10010," "10000," and "10100," which exhibit the smallest means (around 25 kg/m²) and tightest spreads (IQR ≈ 3–4).

- **Cluster 2 (Intermediate-BMI profiles)** groups codes like "00000," "00001," and "00100," centered around the overall sample median (≈ 28 kg/m²) with moderate variability.

- **Cluster 3 (Higher-BMI profiles)** is formed by codes such as "01110," "01111," and "11110," all of which show elevated means (> 30 kg/m²) and larger IQRs (5–7), reflecting greater dispersion.

Because hierarchical clustering incorporates not only central tendency but also distributional spread, it sensitively distinguishes these communities even when mean differences are too subtle (or sample sizes too small) to achieve statistical significance in an ANOVA framework. In practical terms, this suggests that, while individual profile means do not differ enough to pass a formal F-test at n ≈ 2–3 per group, the underlying BMI distributions nonetheless form meaningful low/medium/high strata.

Kruskal Wallis and Dunn's Test Check




```
print(table(df32$cluster3, useNA="ifany"))
```

```
 Low-BMI  Mid-BMI High-BMI
      18        4        10
```

```
>
> # 6. Kruskal–Wallis test on BMI across the three clusters
> kw <- kruskal.test(BMI ~ cluster3, data = df32)
> print(kw)
```

```
	Kruskal-Wallis rank sum test

data:  BMI by cluster3
Kruskal-Wallis chi-squared = 7.4335, df = 2, p-value = 0.02431

>
> # 7. Dunn's post-hoc with Bonferroni adjustment
> dunn <- dunnTest(BMI ~ cluster3, data = df32, method = "bonferroni")
> print(dunn)
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.

          Comparison         Z     P.unadj      P.adj
1 High-BMI - Low-BMI  0.967569 0.333259669 0.99977901
2 High-BMI - Mid-BMI -1.888344 0.058979720 0.17693916
3  Low-BMI - Mid-BMI -2.711384 0.006700294 0.02010088
```

After assigning every one of the 32 samples to Low-, Mid-, or High-BMI clusters (n=18, 4, and 10 respectively), we ran a Kruskal–Wallis test to compare their BMI distributions without assuming normality. The test statistic ($\chi^2 = 7.4335$, df = 2, p = 0.0243) indicates a significant difference in median BMI across the three clusters.

To pinpoint which clusters differ, we performed Dunn's pairwise comparisons with Bonferroni adjustment. Only the Low-BMI vs. Mid-BMI contrast remained significant (Z = –2.7114, unadjusted p = 0.0067, adjusted p = 0.0201), confirming that the Mid-BMI group's BMI distribution is reliably higher than the Low-BMI group's. The High-BMI vs. Mid-BMI comparison (Z = –1.8883, adjusted p = 0.1769) and the High-BMI vs. Low-BMI comparison (Z = 0.9676, adjusted p = 0.9998) did not reach significance after correction.

This pattern differs slightly from our ANOVA/Tukey on aggregate means (which found all three pairwise mean differences significant) because Dunn's test focuses on medians and applies a stringent Bonferroni adjustment—and the Mid-BMI cluster here contains only four observations. Nonetheless, the nonparametric results corroborate the overall Kruskal–Wallis finding that BMI distributions are not identical across clusters, and specifically highlight the distinction between the Low- and Mid-BMI communities. The lack of significance for the High-BMI contrasts likely reflects both the small Mid-BMI sample size and the broader variability in the High-BMI group.

Week 8 conclusion

In Week 8, we explored how body-mass index (BMI) varies across 32 health-profile groups defined by five binary indicators (Sex, HighBP, HighChol, HeartDiseaseorAttack, and Stroke). A classical one-way ANOVA on these groups yielded an overwhelmingly significant result ($F(31, 253\,648) = 438.6$, $p < 2 \times 10^{-16}$) in the full BRFSS dataset, but when restricted to our 32-sample subset (13 observed codes), the ANOVA found no significant mean differences ($F(12, 19) = 1.021$, $p = 0.469$), and Tukey–Kramer pairwise tests likewise produced uniformly large adjusted p-values (all $p > 0.56$). These parametric methods, however, can lose power when group sizes are extremely small (2–3 observations each) and effect sizes modest.

To gain deeper insight, we performed hierarchical clustering on each group's four BMI summary statistics (mean, median, standard deviation, interquartile range). The resulting dendrogram revealed three coherent strata, Low-, Mid-, and High-BMI, characterized by mean values of approximately 25, 28, and over 30 kg/m² and progressively increasing variability. We then overlaid a Gower-based Ward dendrogram onto a heatmap of the five binary risk flags for all 32 samples. This heatmap neatly grouped individuals with few or no flags into the Low-BMI community, those with isolated hypertension (and sometimes high cholesterol) into the Mid-BMI community, and those bearing multiple metabolic and disease flags into the High-BMI community.

Finally, we validated these groupings with nonparametric tests. A Kruskal–Wallis rank-sum test on the three aggregated clusters confirmed a significant difference in median BMI ($\chi^2 = 7.43$, $df = 2$, $p = 0.024$), and Dunn's post-hoc comparisons (Bonferroni-adjusted) highlighted a robust median-BMI gap between Low- and Mid-BMI clusters (adjusted $p = 0.020$). Although contrasts involving the High-BMI cluster did not reach significance, likely due to its greater within-group variability and the small size of the Mid-BMI group, the overall pattern corroborates that accumulating health flags tracks closely with increasing BMI. Together, these complementary methods, ANOVA/Tukey, HC-tree clustering, and nonparametric testing, provide a comprehensive, visually intuitive, and statistically rigorous stratification of BMI risk.