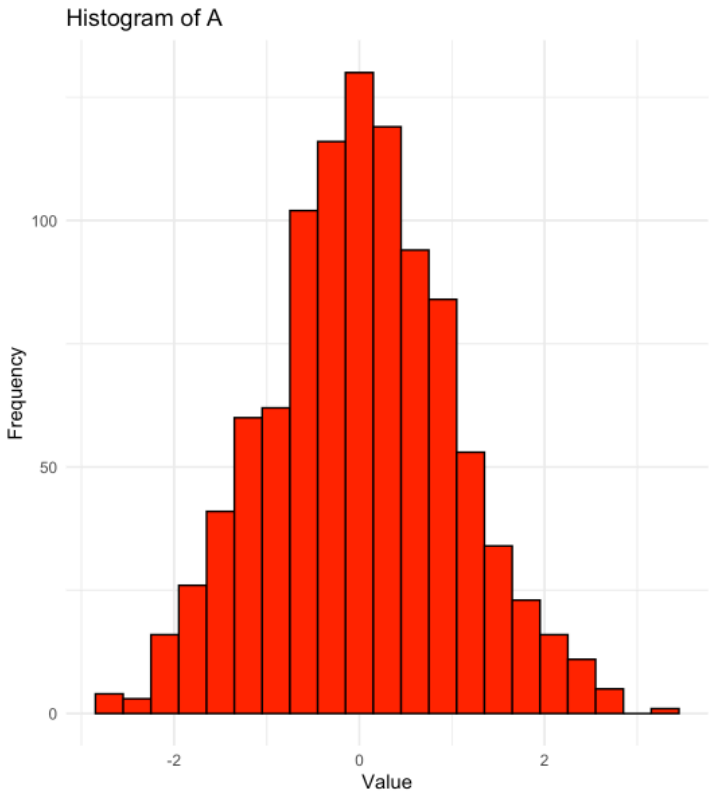
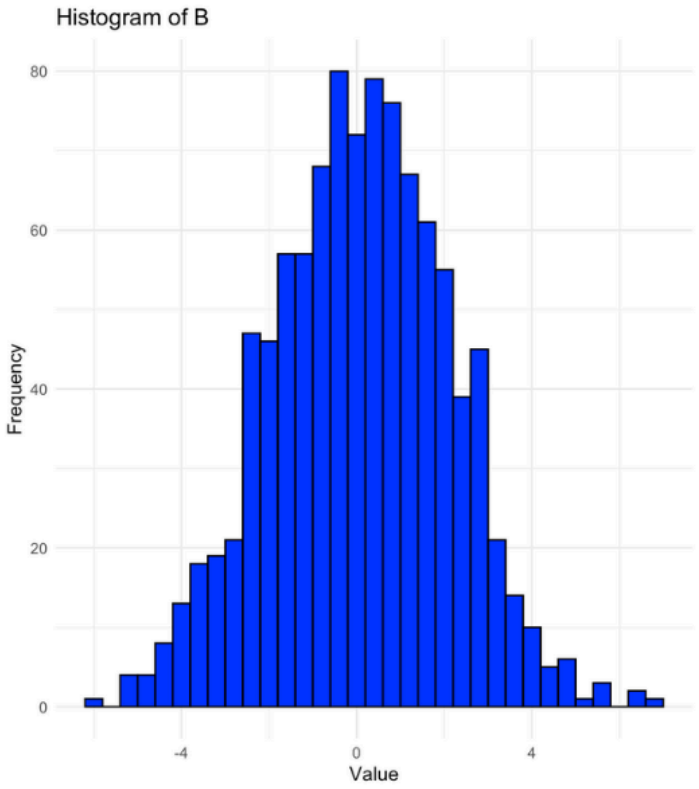
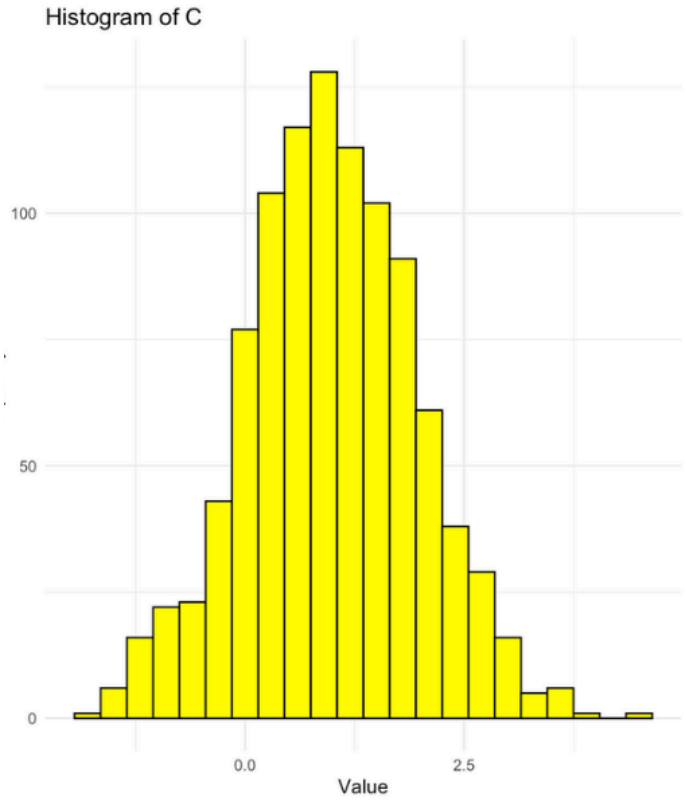
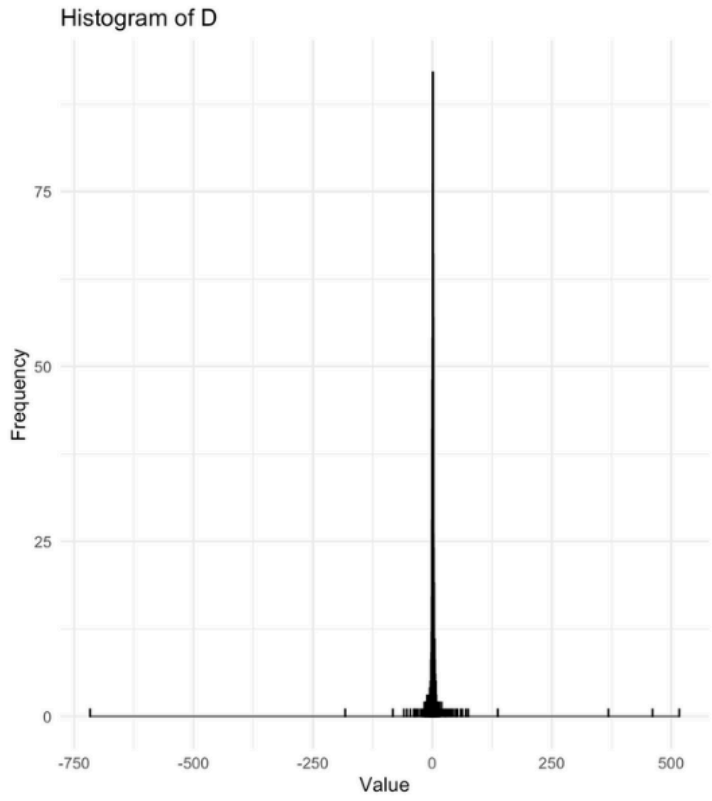
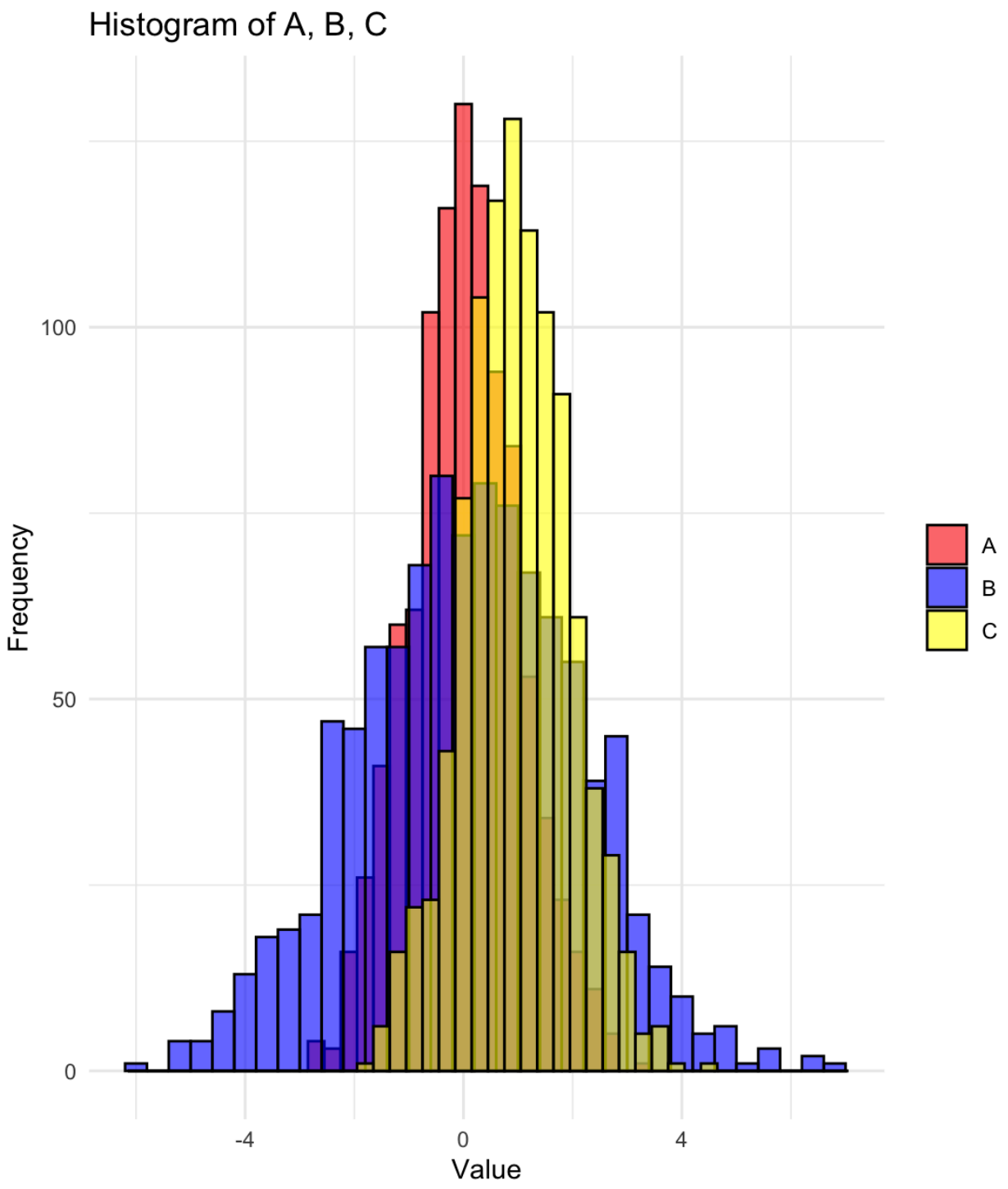
Week 1 (Histograms)
Graphs

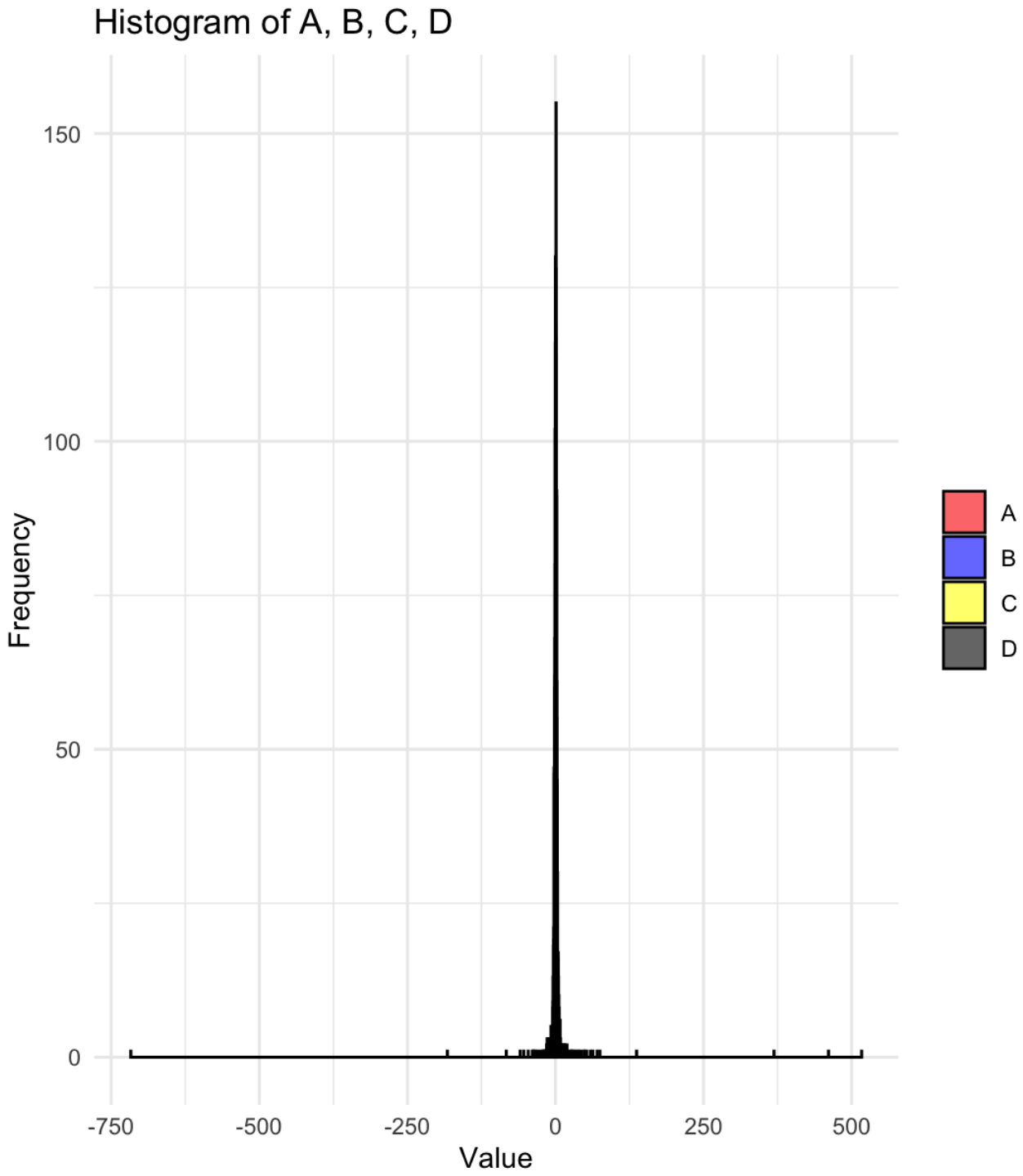









Histogram of A, B, C
Frequency
100
50
0
-4
0
4
Value
A
B
C


Histogram of A, B, C, D
150
100
50
0
Frequency
-750
-500
-250
0
250
500
Value
A
B
C
D

Analysis

Variable A

The histogram for **A** is a tall, symmetric mound centred almost exactly at 0. Both mean (≈ 0.02) and median (≈ 0.01) lie on the peak, and the spread is modest (s ≈ 1). The data taper smoothly to the left and right, with virtually all observations contained in the -3 to +3 range. Only a handful of points sit in the outermost bins, so there is no evidence of problematic outliers or skewness. Visually—and confirmed by the numerical summary—A behaves like a textbook Normal sample, making it a reliable candidate for any parametric procedures that assume normality and equal variance.

Variable B

B is also unimodal and roughly centred near 0, but its bell is noticeably wider (sample standard deviation ≈ 2). Extreme values extend to about -6 and +7, giving the distribution heavier tails than a perfect Normal. The shape remains largely symmetric—mean and median are almost identical—but the extra spread implies larger standard errors whenever B enters a model. Mild tail heaviness is not a fatal flaw; however, keep in mind that inference based on Normal theory may be slightly liberal unless we use a large-sample adjustment or a robust alternative.

Variable C

The histogram for C is the narrowest of the three “regular” variables. It forms a steep, nearly symmetric peak around 1, with most data between −0.5 and 3.5. Dispersion (sample standard deviation ≈ 1) is similar to A even though the centre is shifted to the right. There is a faint suggestion of right skew, but both tails fall away quickly and no single bar sticks out as an outlier. Because C is tight and approximately Normal, it will contribute the most precise estimates in any combined analysis.

Variable D

D is in a league of its own. The histogram shows an extreme, needle-like spike at 0 flanked by very sparse bars that stretch out past ±500. Skewness is strongly negative (longer left tail), and kurtosis is colossal. This pattern indicates a mixture of “ordinary” values crowded near 0 and a scattering of extreme shocks. The variance of D (sample standard deviation ≈ 35) dwarfs that of A–C, so including D untransformed will dominate any scale-sensitive statistics (means, sums of

squares, etc.) and violate normality and homoscedasticity assumptions. Practical options are to analyse D separately, Winsorise or trim the tails, or switch to methods that rely only on order information (e.g., rank tests).

Overlay of A, B, C

Plotted together, the three distributions share a common centre near 0 but differ in spread:

- **C** (yellow) is the tight core.
- **A** (red) surrounds C with moderately wider shoulders.
- **B** (blue) has the broadest base and the heaviest tails.

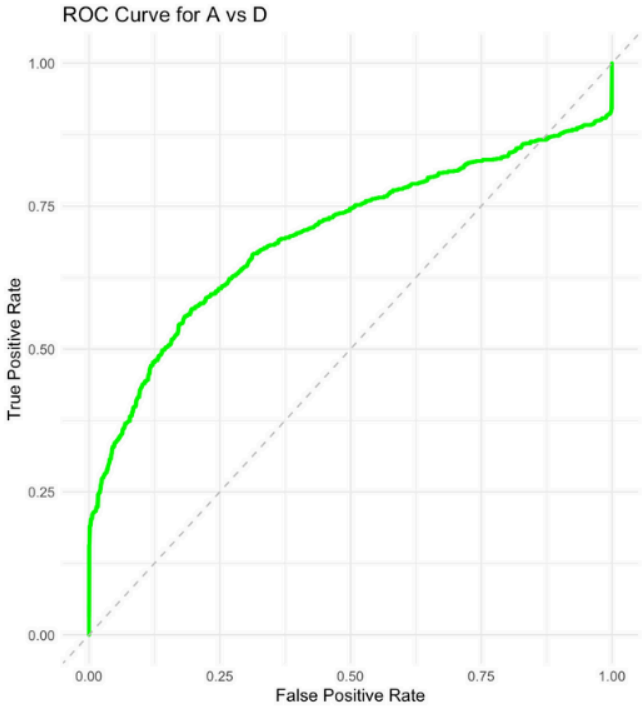
Despite these differences, all three are unimodal and roughly symmetric, so they plausibly arise from the same underlying family with different variances. If a later analysis pools them, we should either verify the equal-variance assumption formally or adopt a method (e.g., Welch's t-test) that allows variances to differ.

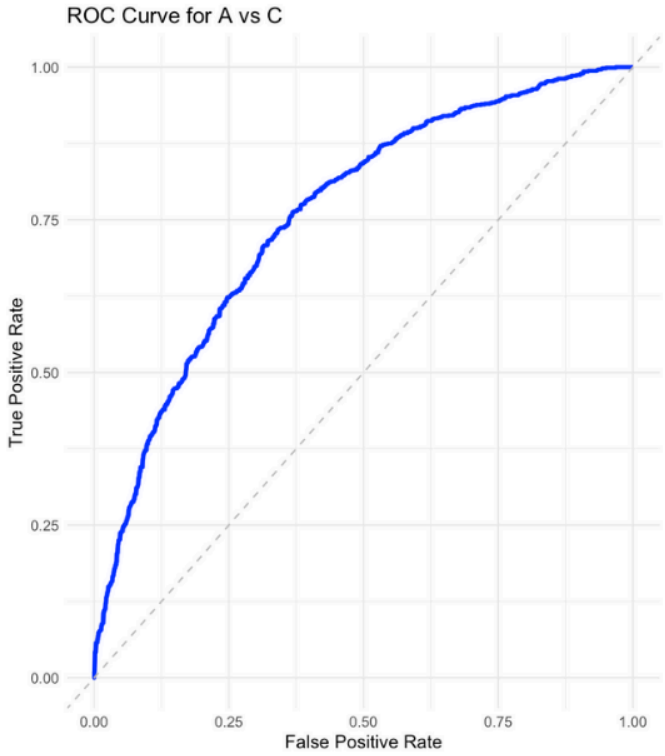
Overlay of A, B, C, D

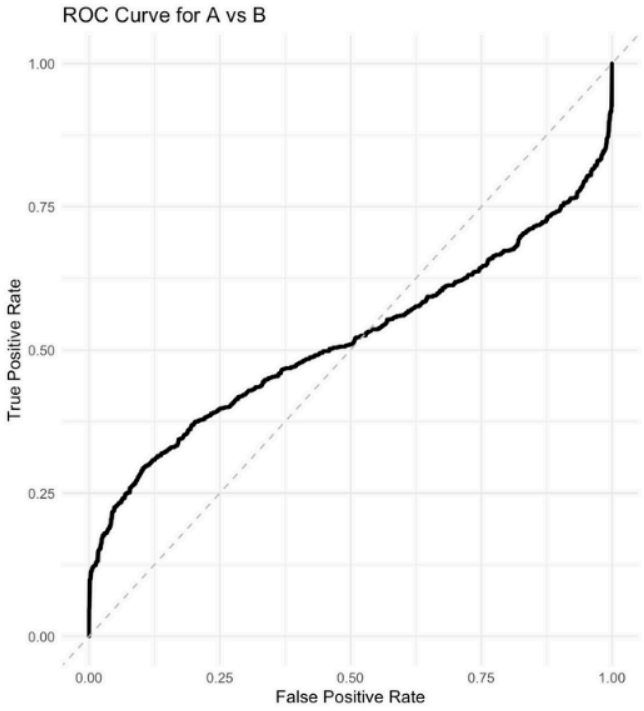
Adding D compresses the x-axis so severely that the structure of A–C becomes invisible. This illustrates how a single high-variance variable can obscure the rest of the data and distort graphical and numerical summaries. Any combined analysis should therefore exclude D or transform it to a comparable scale before drawing joint conclusions.

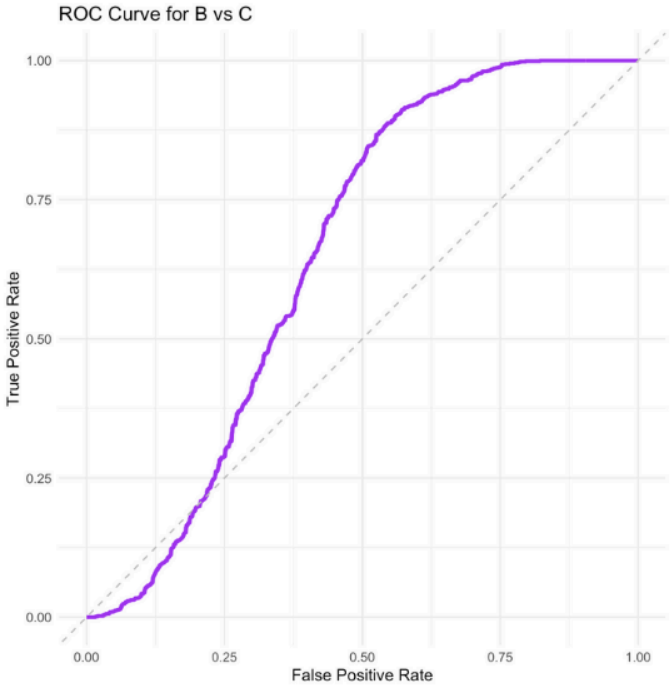
In summary, visual inspection of the histograms shows that variables A, B, and C follow roughly Normal patterns with varying spreads, while D is a clear outlier in both shape and scale. This initial analysis sets the stage for more formal checks like ROC curves and parametric testing.

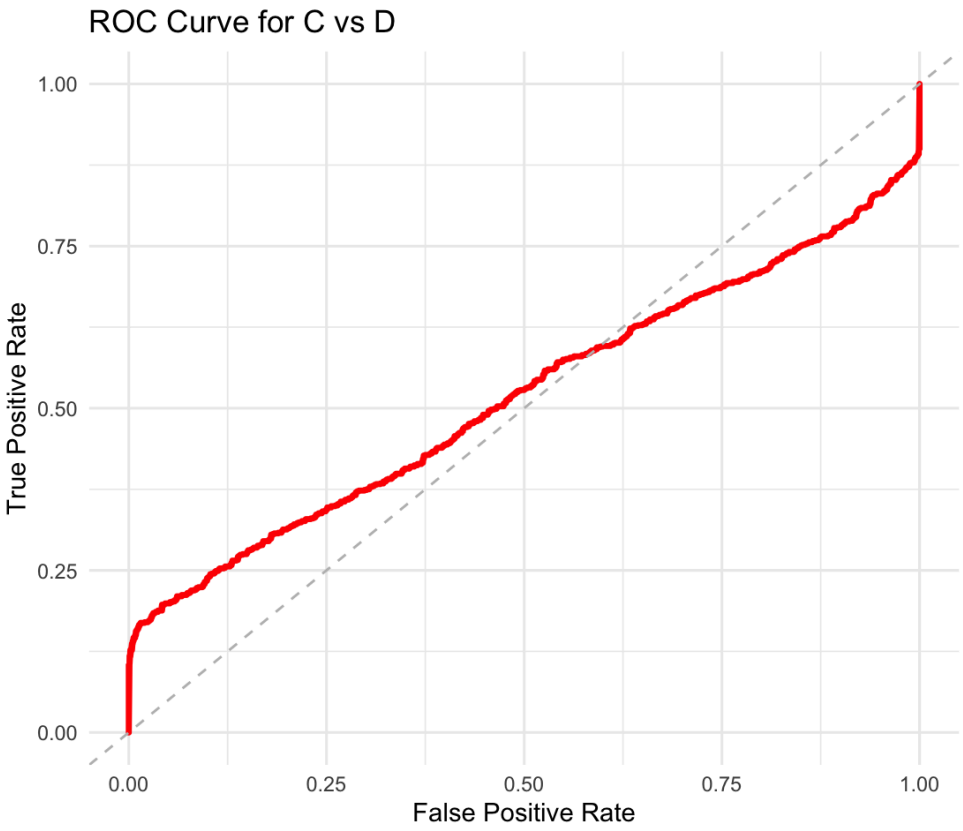
Week 2 (ROS curves)
Graphs

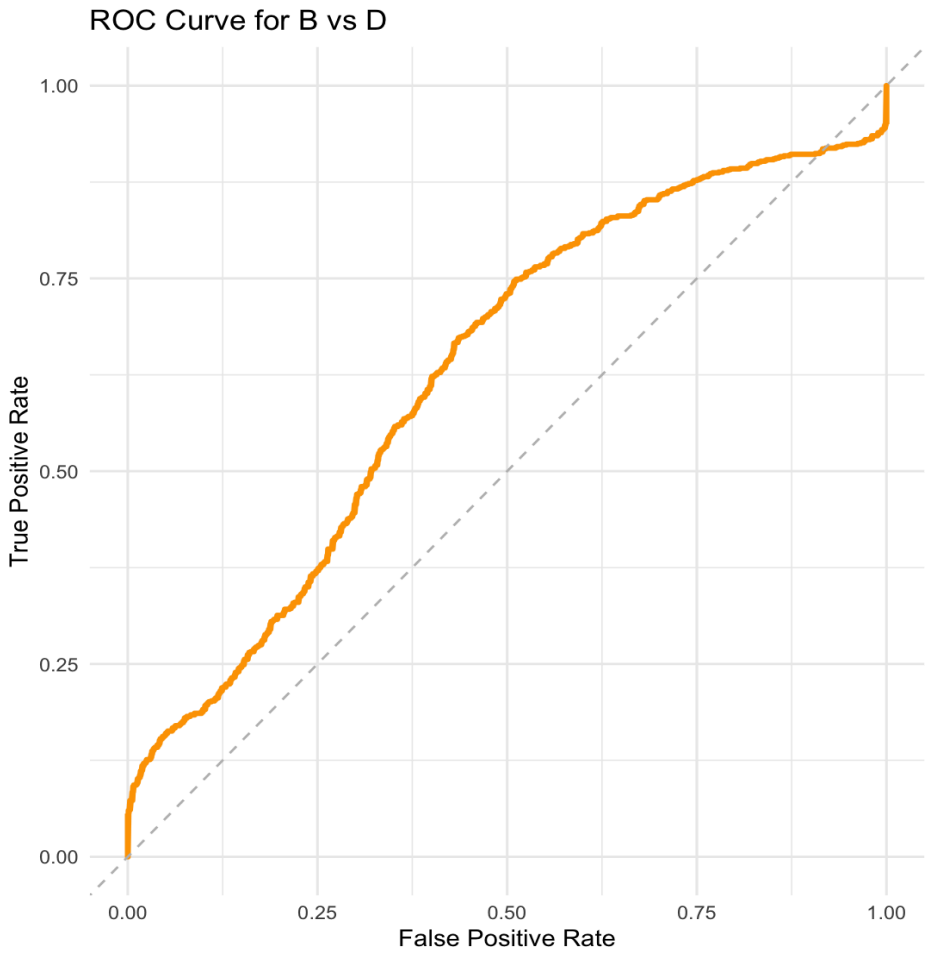


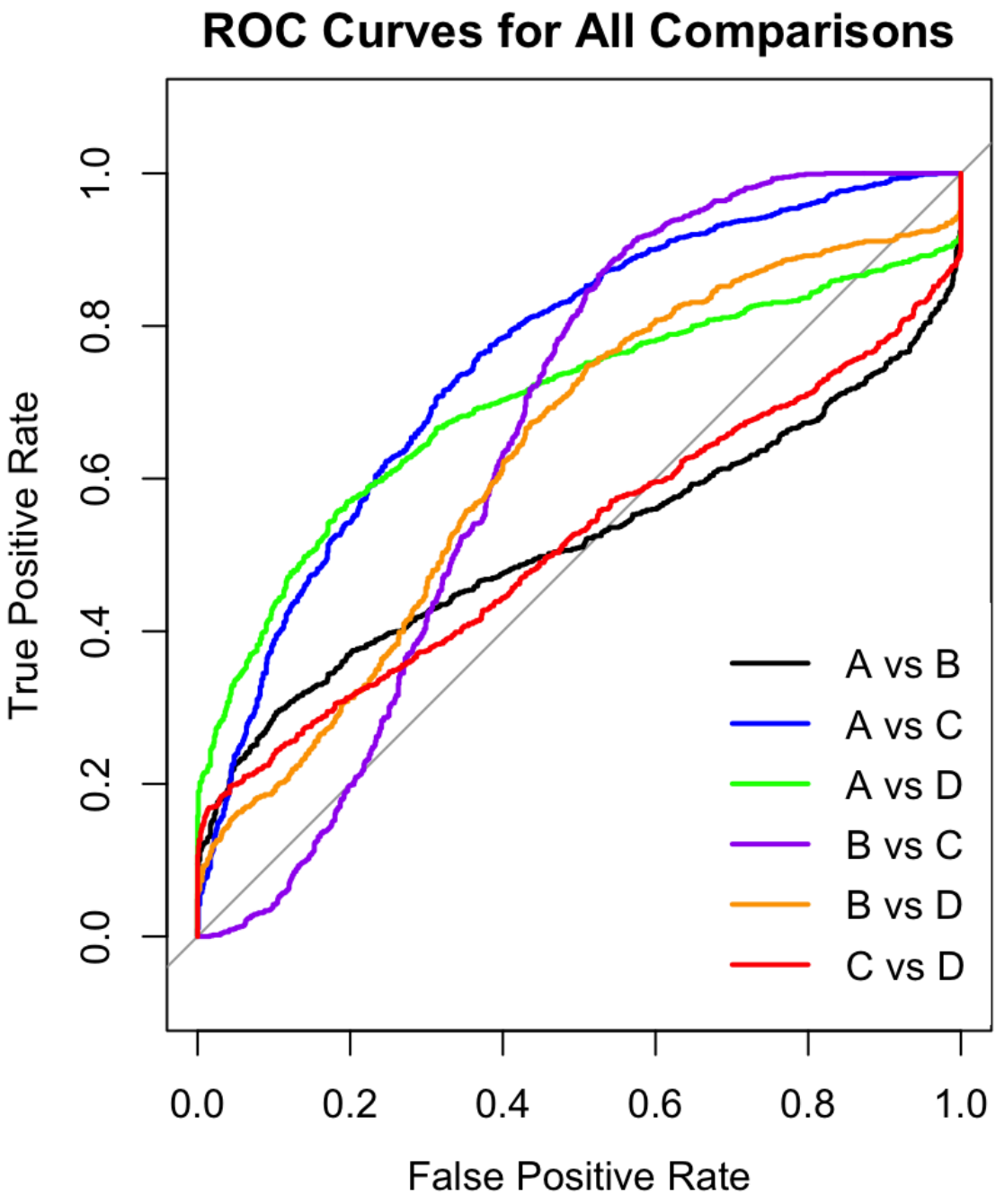







ROC Curve for C vs D
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00


ROC Curve for B vs D
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00


ROC Curves for All Comparisons
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
A vs B
A vs C
A vs D
B vs C
B vs D
C vs D

ANALYSIS

The receiver operating characteristic (ROC) framework allows us to evaluate how well a single continuous variable can distinguish between two labeled groups. For each pair of distributions (A, B, C, D), I plotted the ROC curve and calculated the Area Under the Curve (AUC). The AUC provides a numerical measure of discriminative ability:

- AUC < 0.60 → Chance level
- 0.60–0.70 → Poor
- 0.70–0.80 → Fair
- 0.80–0.90 → Good
- AUC > 0.90 → Excellent

Here are the AUC values for each pairwise comparison:

- **A vs B**: 0.5163
- **C vs D**: 0.5171
- **A vs D**: 0.698
- **B vs C**: 0.650
- **B vs D**: 0.6297
- **A vs C**: 0.7571

A vs C (AUC = 0.757) – Best Separator

The ROC curve for A versus C rises steeply from the origin, reaching a true positive rate (TPR) of approximately 60% while keeping the false positive rate (FPR) below 20%. This strong performance aligns with the Week 1 histograms: variables A and C have similar spreads (standard deviation ≈ 1), but their means differ by about one unit. Because their distributions only overlap at the edges, this comparison offers the clearest separation. A threshold around FPR ≈ 0.18 captures about 58% of class C while misclassifying very few values from class A. This makes A vs C the most reliable stand-alone classifier.

A vs D (AUC = 0.698) – Fair but Limited

The ROC curve for A versus D rises above the diagonal line but less sharply than in the A vs C comparison. Based on Week 1, we know that variable D is a mixture of a sharp spike at 0 and extremely long tails. The spike overlaps heavily with A's center, limiting

discrimination. However, the extreme values in D's tails provide some useful signal, allowing the AUC to approach 0.70. A threshold at FPR ≈ 0.22 gives a TPR of about 60%, which may be acceptable if false positives are tolerable in exchange for catching more true D cases.

B vs C (AUC = 0.650) – Weak to Fair

The ROC curve for B versus C follows a gradual S-shape. To achieve a TPR of 70%, we would need to accept an FPR of about 45%. This lower discriminative power can be explained by the Week 1 histograms: B has a much wider spread than C, so many B values overlap with the core of C's distribution. Although B vs C may not be strong enough for use alone, it could still be helpful in a multivariable model when combined with other predictors.

B vs D (AUC = 0.630) – Poor Separation

The ROC curve for B versus D stays close to the diagonal, showing that these two variables have considerable overlap. B's wide distribution overlaps with both the central spike and tails of D, leaving very little space to separate the two classes. Without transformation or additional features, this comparison contributes little useful information and should be given low priority in any modeling strategy.

A vs B (AUC = 0.516) & C vs D (AUC = 0.517) – No Discriminative Power

Both of these ROC curves fall nearly on the diagonal, indicating performance no better than random guessing. The Week 1 histograms explain this result. A and B have nearly identical centers, and B's wider spread means their distributions heavily overlap. Similarly, C and D are both centered at the same value: C's narrow bell sits right on top of D's spike. Because of this complete central overlap, neither comparison provides useful separation. These variable pairs should be excluded from classification tasks unless further transformed.

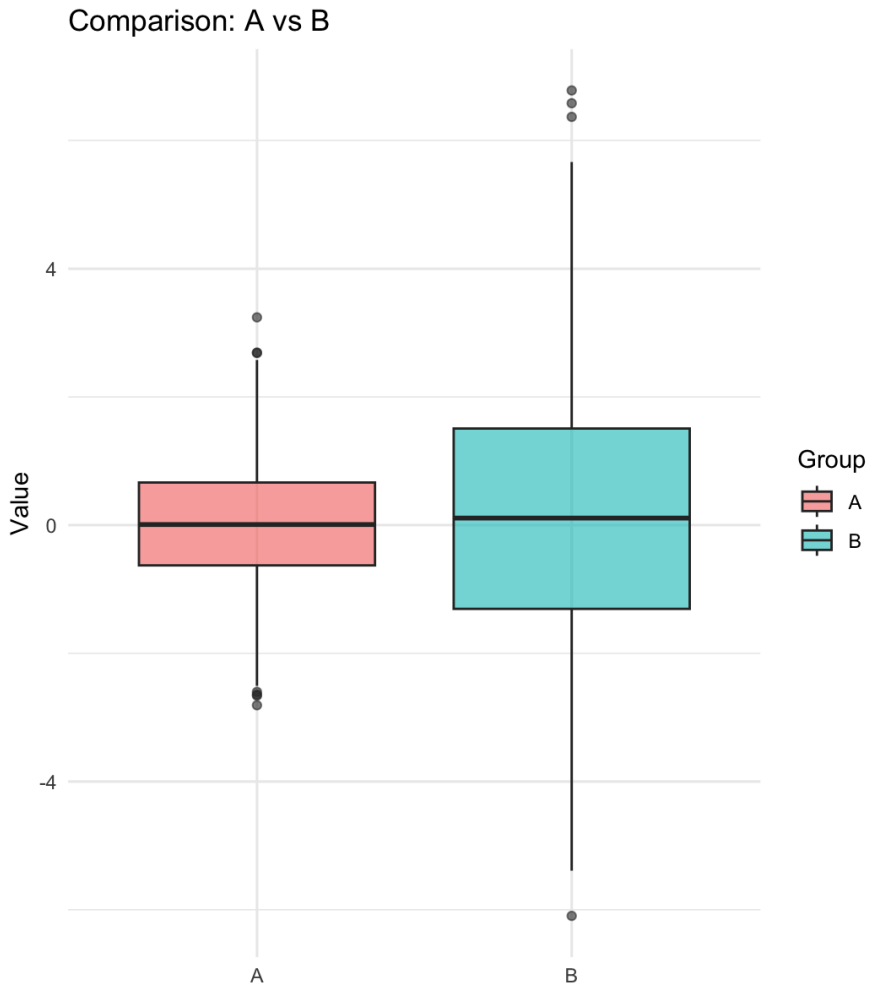
Summary and Connection to Week 1

The ROC curve results strongly reflect what we observed in the Week 1 histogram analysis:

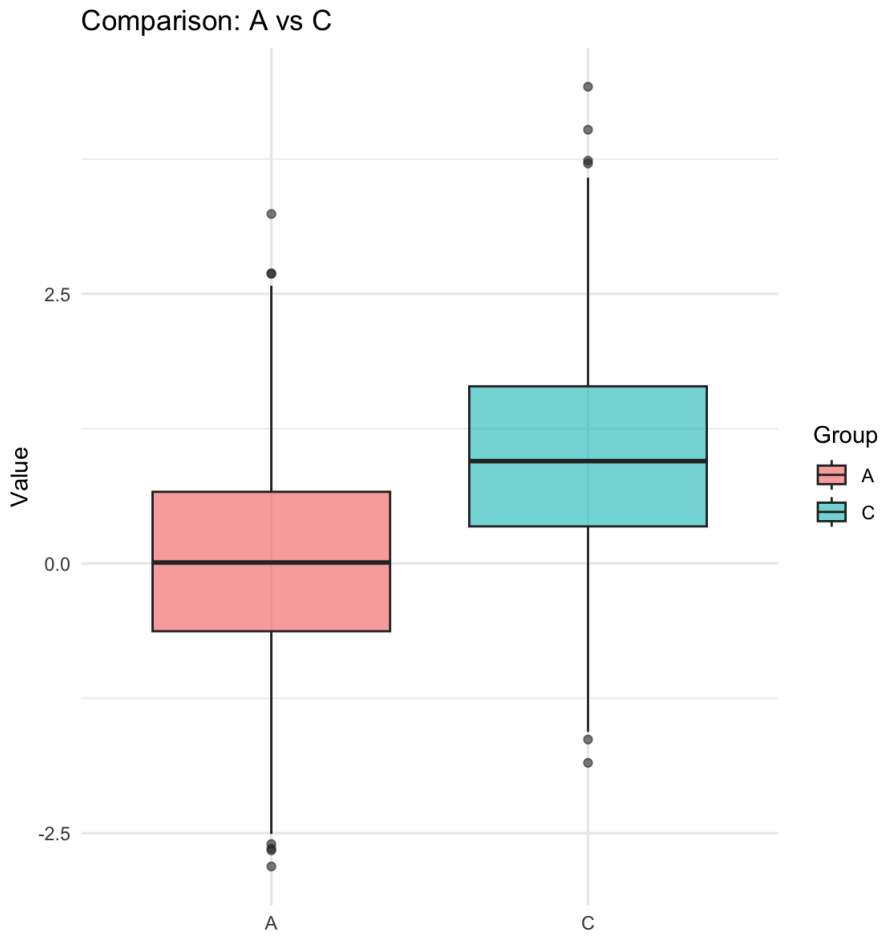
- **Clear separation of means with similar spreads** results in high AUC values. This is seen in **A vs C**, where the distributions are centered apart but have equal variance.
- **High variance without much mean shift** leads to weaker discrimination. This was evident in all comparisons involving **B**, which had the widest spread of all four variables.
- **Distributions with a sharp central spike and heavy tails**, like **D**, offer moderate classification ability. In **A vs D**, the tails provide a useful signal, but the central overlap with A limits the overall AUC.
- **Heavy central overlap between distributions** results in AUCs close to 0.50. This was the case in **A vs B** and **C vs D**, where the distributions aligned closely at their centers.

These results confirm that our initial visual analysis using histograms was an effective predictor of formal ROC performance. The relationships between center, spread, and shape that we observed in Week 1 directly explain the patterns we now see in ROC-based classification.

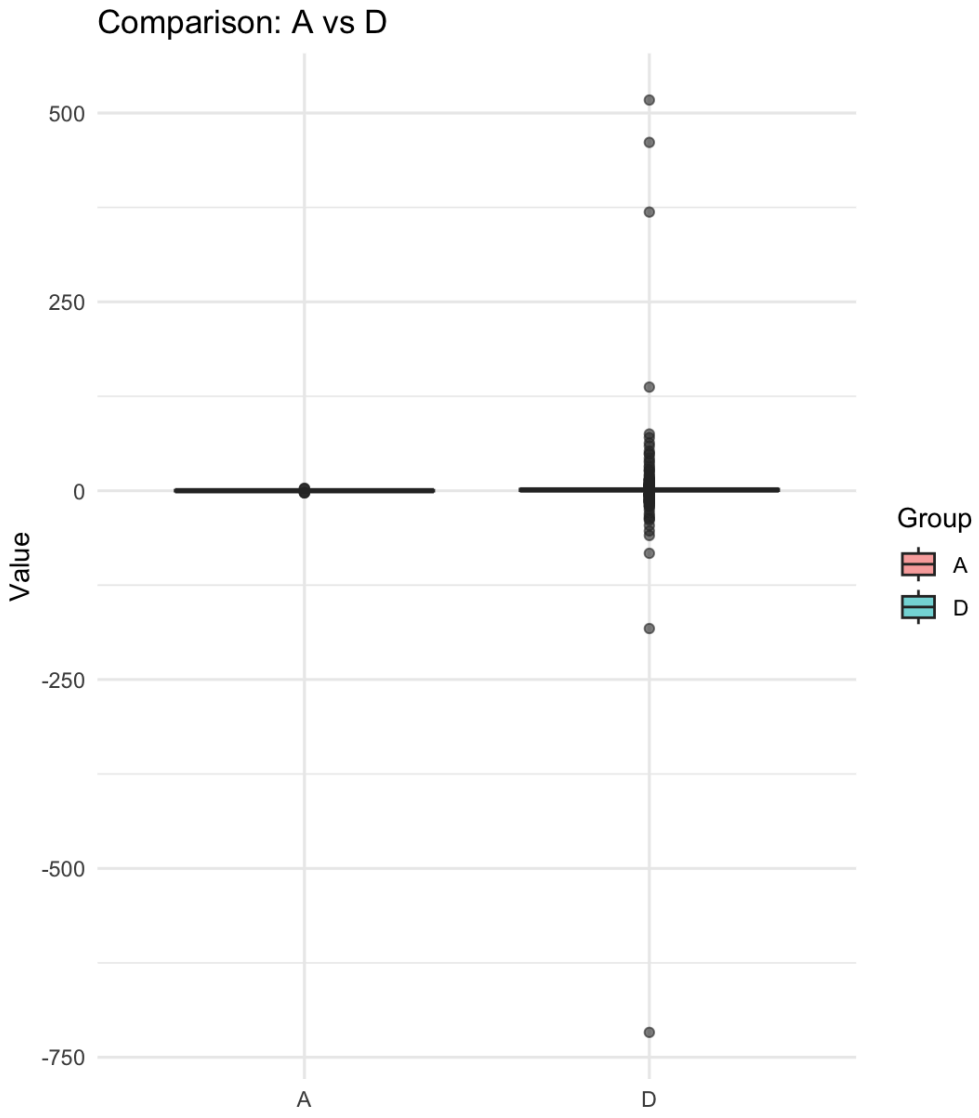
Week 3 (Two Sample T test)



The t-test comparing groups A and B resulted in a p-value of **0.3336** and a 95% confidence interval of **[−0.208, 0.071]**, indicating no statistically significant difference between their means. The estimated means (A ≈ 0.016, B ≈ 0.085) are quite close, and the distributions overlapped almost entirely in the histogram. This explains the high p-value and the low discriminative ability observed in the ROC curve (AUC ≈ 0.52).This aligns with both the histogram (Week 1), where A and B largely overlapped. There is no evidence that these two groups differ in a meaningful way based on their sample means.

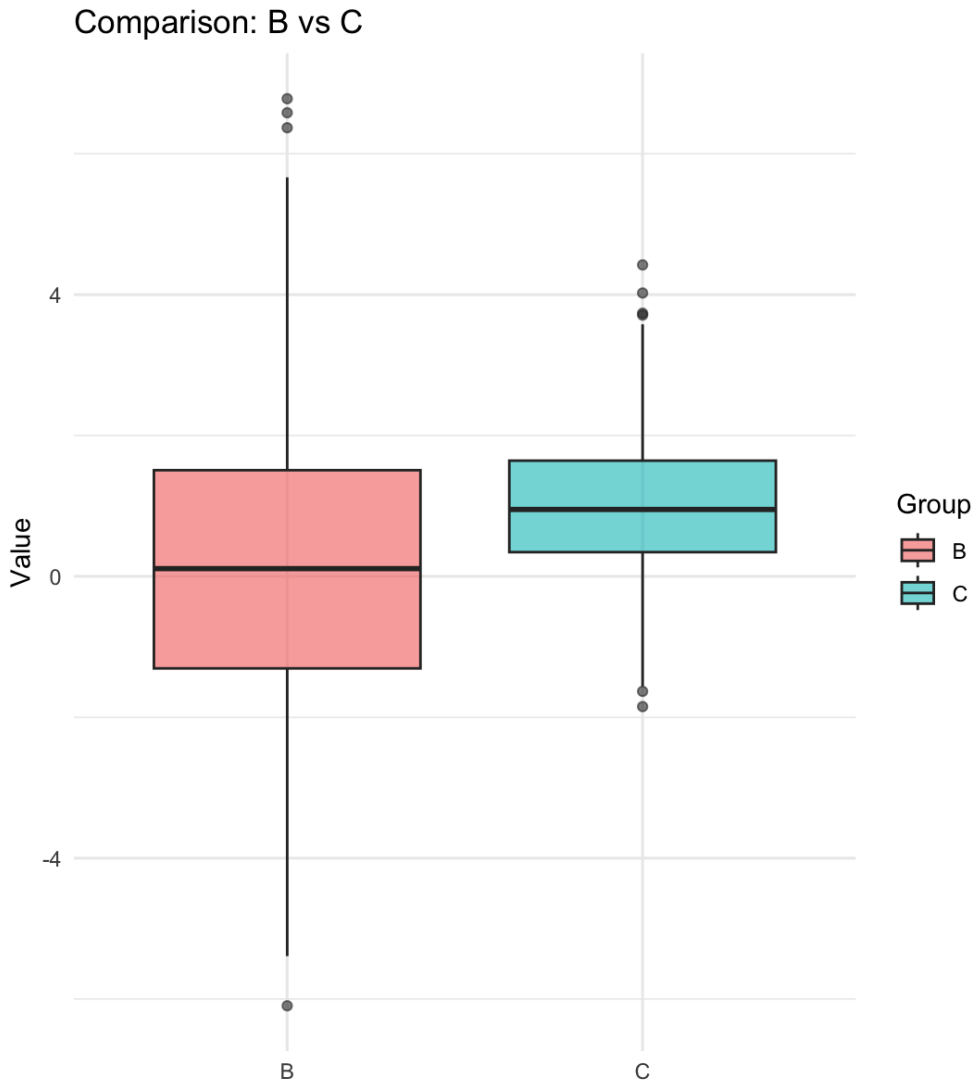


This comparison showed a highly significant p-value of **< 2.2e-16**, with a confidence interval of **[−1.050, −0.877]**. The difference in sample means is clear (C ≈ 0.98 vs A ≈ 0.016), and the histograms supported this with a visible shift in center. This large effect size and the small p-value indicate a strong difference between the two groups. The ROC curve (AUC ≈ 0.76) also confirmed strong separation. This is a clear and reliable comparison with strong evidence for a true difference in population means.



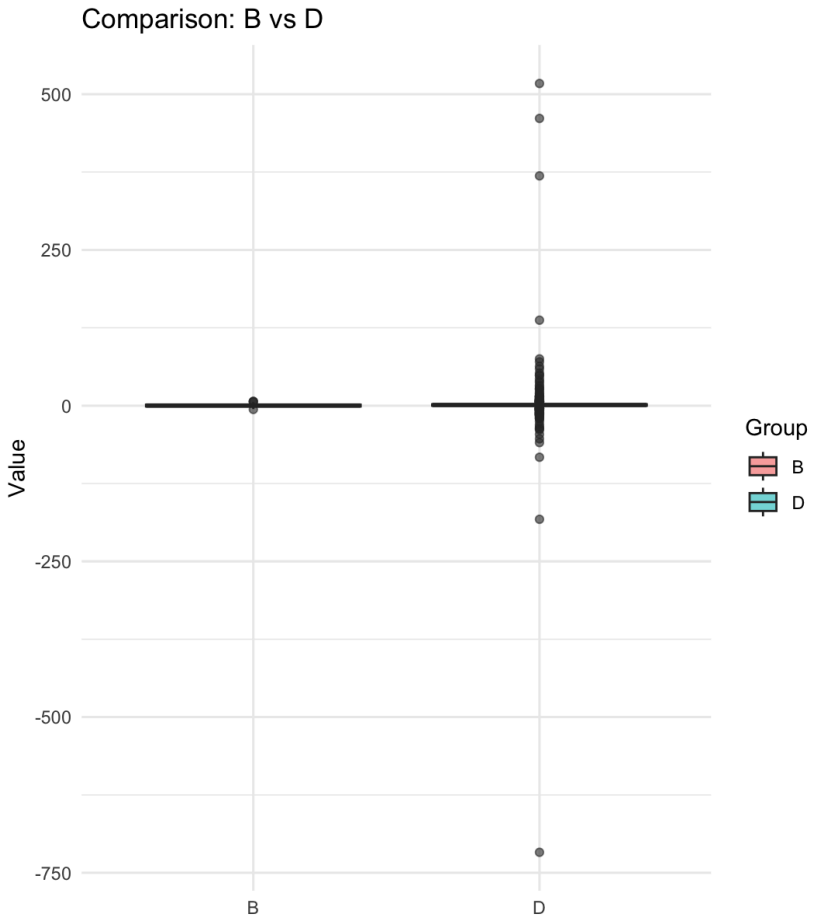
Although the sample means differ (D ≈ 1.72 vs A ≈ 0.016), the p-value is **0.1283**, and the confidence interval **[−3.90, 0.49]** includes zero, meaning the test is not statistically significant. This result is affected by the high variance in group D, which showed both a tight spike and extreme outliers. The large spread inflates the standard error and weakens the test's ability

to detect a difference, even if one exists. While the mean difference seems large, it's not reliable under these conditions.

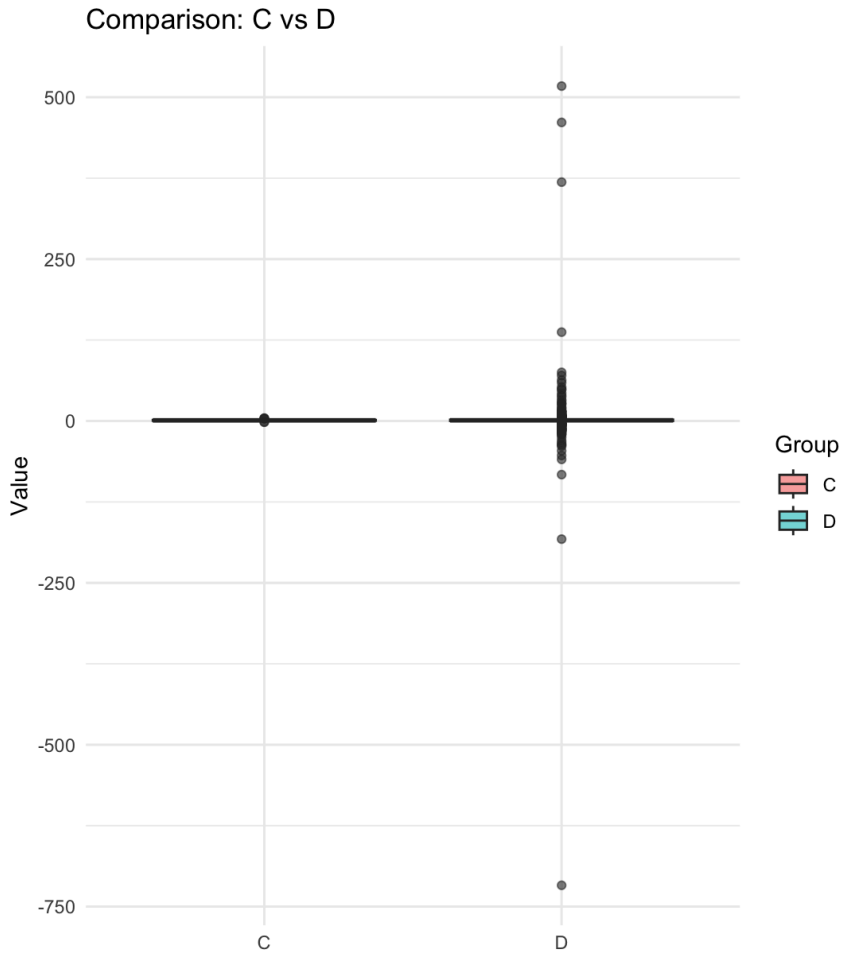


The B vs C comparison yields a p-value of **< 2.2e-16** and a confidence interval of **[−1.034, −0.756]**, indicating a strong and statistically significant difference in means. Although group B has a wider spread, the center of group C is clearly shifted. This is supported by the histogram and by the ROC curve (AUC ≈ 0.65), which showed decent separation. Even with

unequal variances, the difference in means is strong enough to be confidently detected.

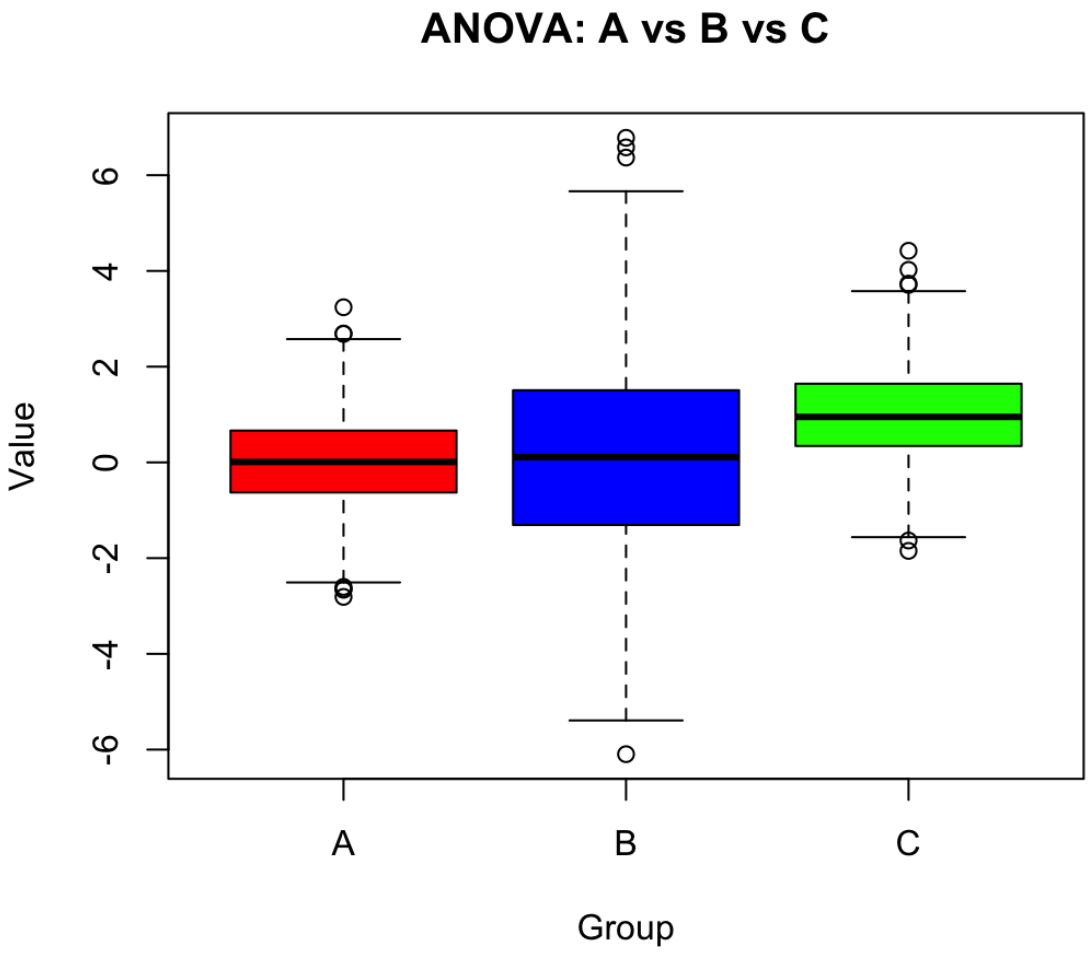


This test resulted in a p-value of **0.1449** and a confidence interval of **[−3.84, 0.56]**, which is not statistically significant. Like the A vs D comparison, this result is likely due to the extreme variability in group D. Despite a noticeable gap in means (D ≈ 1.72 vs B ≈ 0.085), the test lacks power due to the large spread in D's distribution. The ROC curve also suggested weak separation (AUC ≈ 0.63), reinforcing that this comparison is inconclusive.



The final comparison between C and D also showed a non-significant result, with a p-value of **0.5083** and a wide confidence interval of **[−2.94, 1.46]**. Although C is tightly centered and D has a higher average, the extreme variability in D's distribution makes it difficult to detect a stable difference in means. The overlap and inconsistency in D's shape reduce the reliability of this test result. The corresponding ROC curve (AUC ≈ 0.52) showed near-random performance, which aligns with this outcome.

Week 4 (One-Way ANOVA)



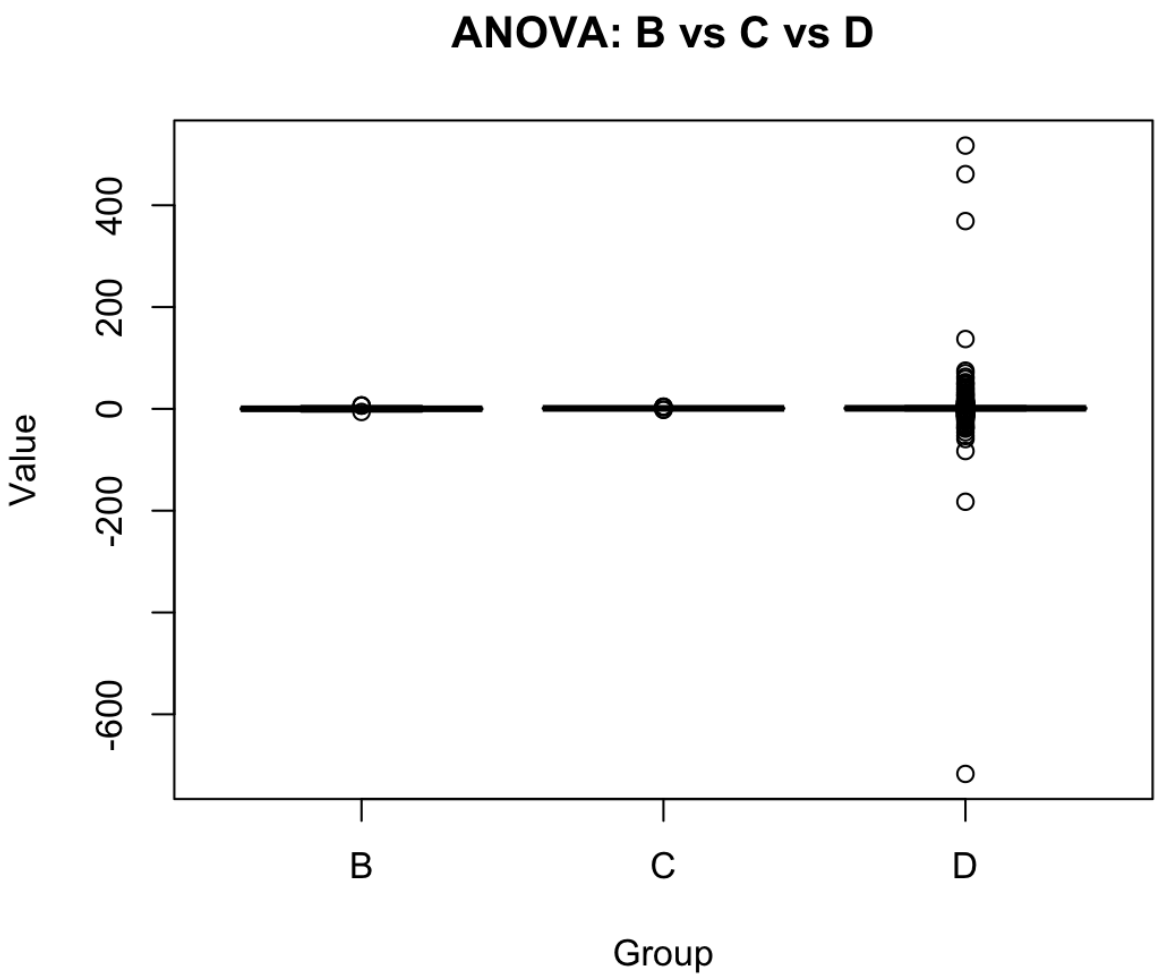
A one-way ANOVA was conducted to test whether the mean values of groups A, B, and C were significantly different. The test returned an F-statistic of 144.1 and a p-value < 2e−16, which is highly significant. This indicates strong evidence that at least one of the three group means differs from the others.

This result aligns perfectly with the patterns we observed in previous weeks:

- In Week 1, the histograms revealed that group A was centered around 0 with moderate spread, group B had a similar center but a much wider distribution, and group C had the most compact shape but was centered further to the right (around 1). These differences in both spread and center are reflected in the clear separation of the boxplots, especially the elevated position of C.

- In Week 2, the ROC curve for A vs C had the highest AUC (~0.76), indicating strong classification power, and the curve for B vs C also showed fair separation. These ROC results hinted at the kind of differences ANOVA has now confirmed.

A follow-up Tukey HSD test showed that C differs significantly from both A and B, while the comparison between A and B was not statistically significant — consistent with their overlapping shapes and centers in earlier weeks. This confirms that group C's upward shift is the most meaningful difference in this trio, while A and B share a common central tendency despite differing variances.



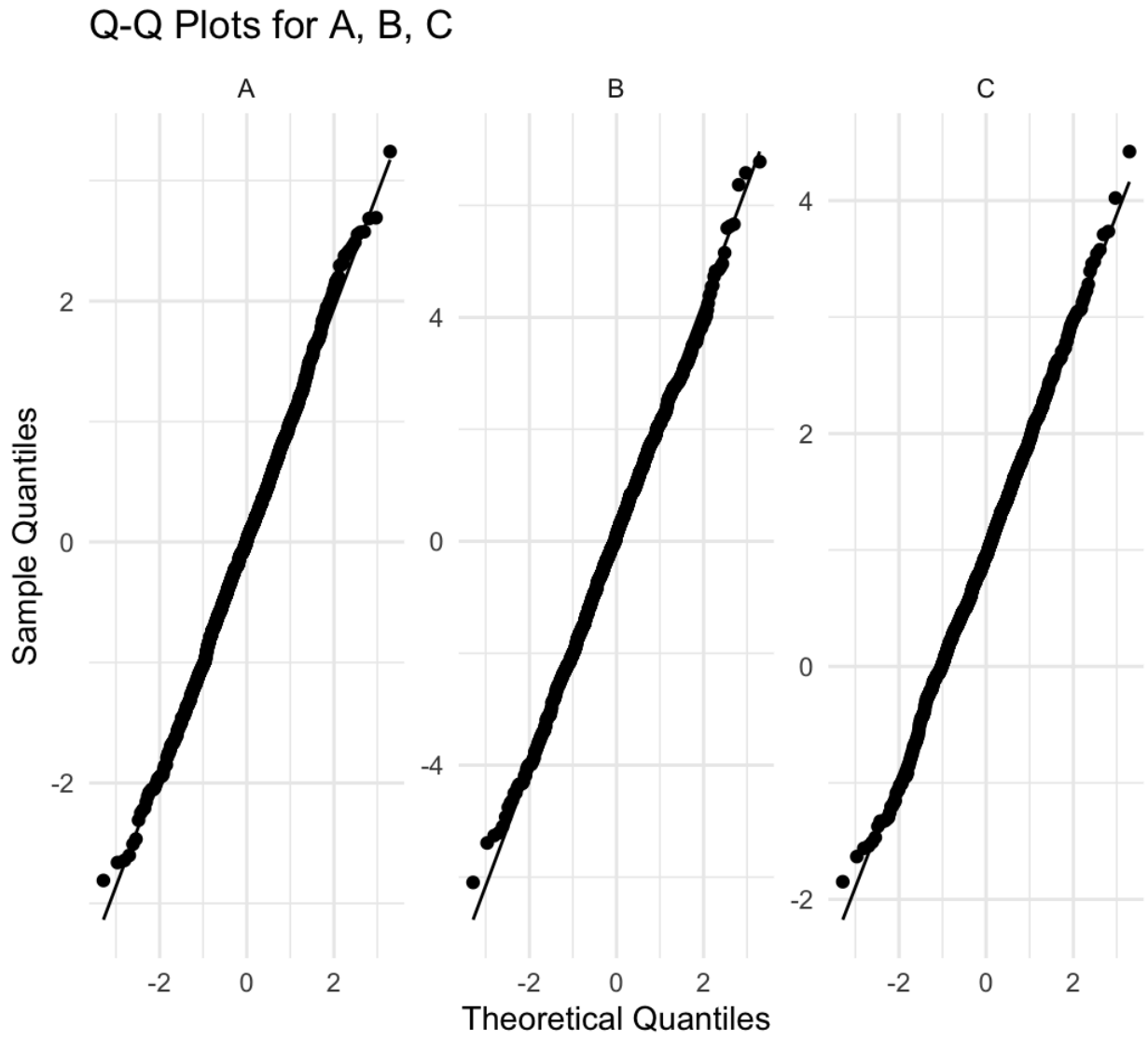
The second ANOVA, comparing groups B, C, and D, returned an F-statistic of 1.6 with a p-value of 0.202, which is not statistically significant. This means there is no reliable evidence that the mean values differ between these three groups.

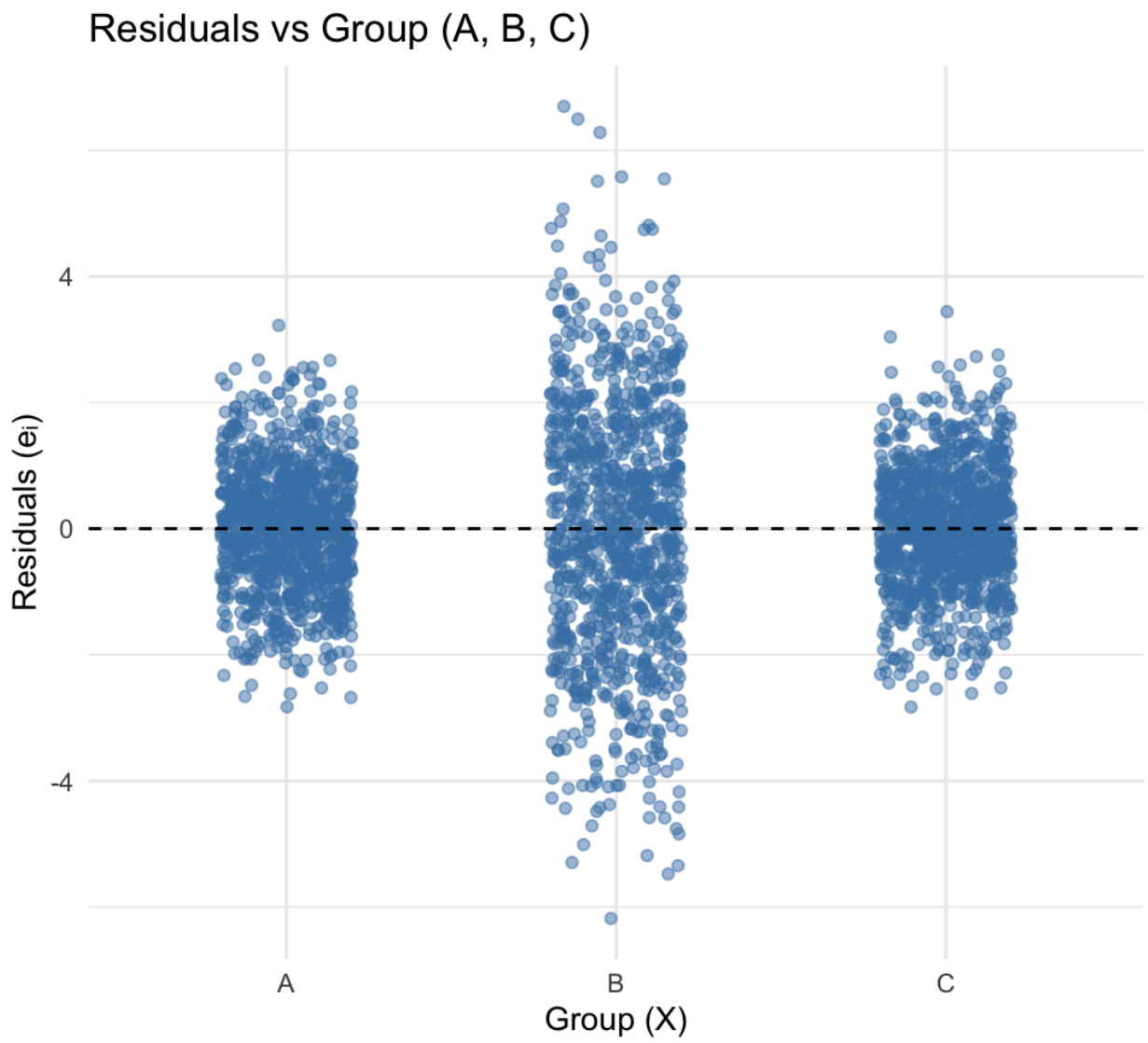
This outcome is supported by what we saw earlier:

- In Week 1, group D had an extreme shape — a tight spike near 0 combined with massive tails. While its variance was huge (making it visually striking), its central location overlapped both B and C, which contributed to the high overlap and lack of mean shift.
- In Week 2, the ROC curves for B vs D and C vs D both had AUCs close to 0.5, indicating weak classification — and again, suggesting that these groups cannot be reliably separated by a simple mean comparison.
- The boxplot from Week 4 confirms this, showing that although D has wide spread, the medians of all three groups are close, and the large variance in D adds noise rather than signal. This explains why the ANOVA failed to find a significant difference.

As the ANOVA was not significant, no Tukey post-hoc test was needed for this comparison.

Week 5 (ANOVA Assumption with 95% confidence)





ABC:

To evaluate whether variables A, B, and C meet the assumptions of ANOVA, we conducted Shapiro-Wilk normality tests, variance comparisons, and graphical diagnostics (Q-Q plots and residual plots).

Normality Check:

The Shapiro-Wilk p-values were:

- A: p = 0.4765
- B: p = 0.6316

- C: p = 0.5889

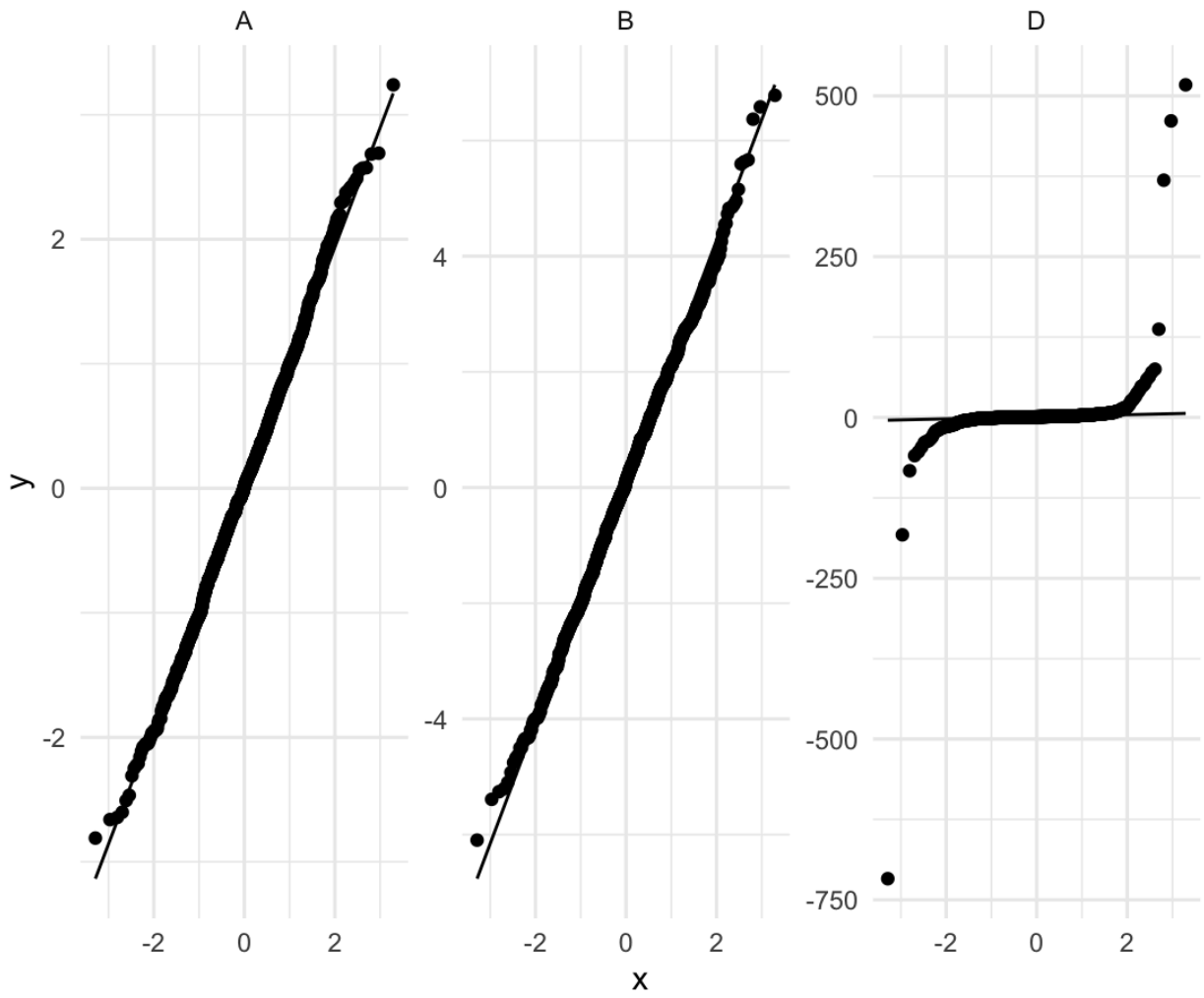
All p-values are greater than 0.05, so we fail to reject the null hypothesis of normality. This suggests the values for all three groups are approximately normally distributed. This conclusion is visually supported by the Q-Q plots, which show that the sample quantiles follow the theoretical Normal line fairly well with only mild deviations at the tails.

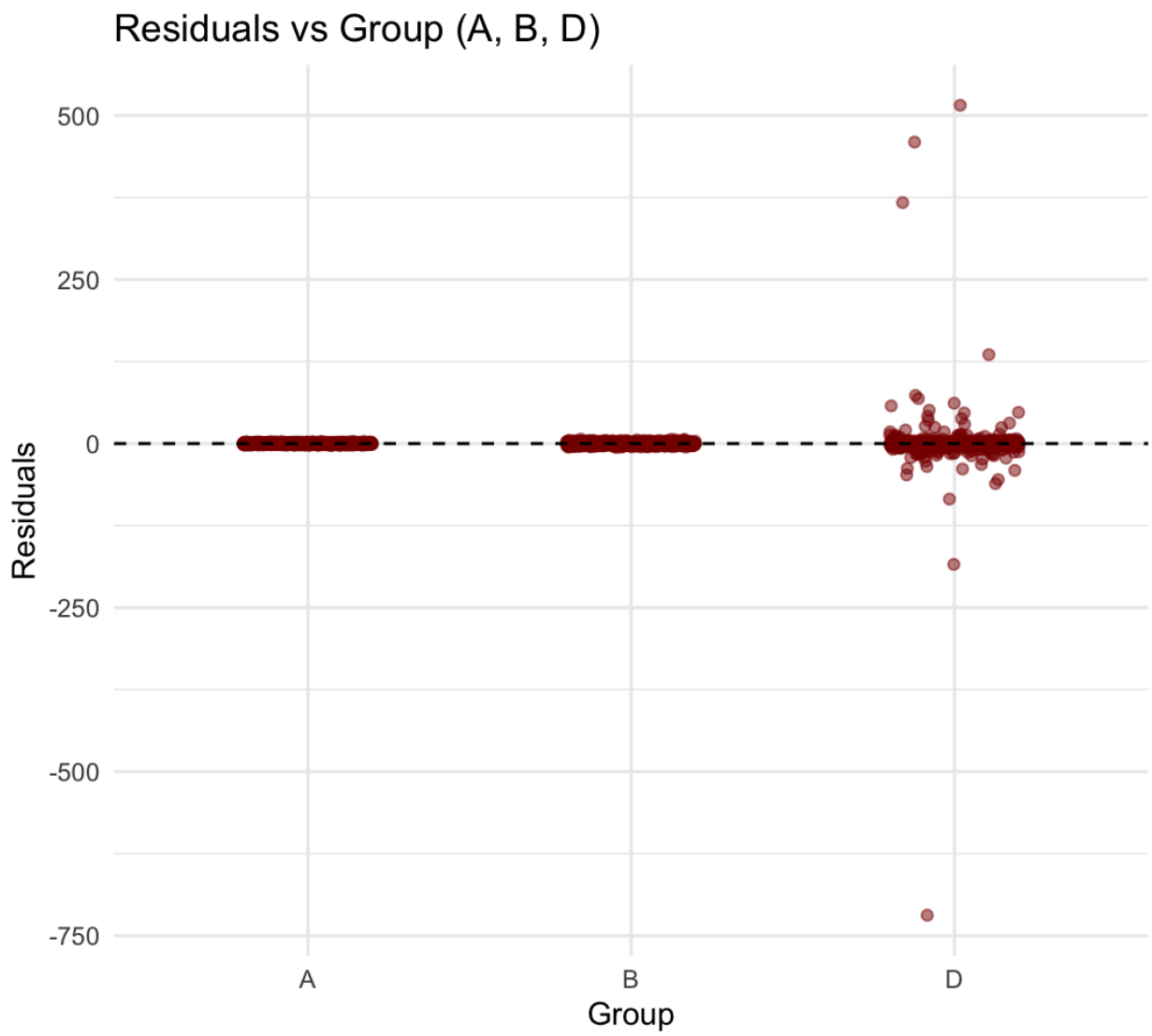
Equal Variance Check:

Sample variances were:

- var(A) = 0.98
- var(B) = 4.08
- var(C) = 0.96

The max-to-min variance ratio is 4.26, which is slightly above the typical threshold of 4. This suggests that the equal variance assumption may be questionable, especially due to B having a noticeably larger spread. The residual plot supports this: residuals for group B appear more dispersed compared to A and C, hinting at mild heteroscedasticity.


Q-Q Plots for A, B, D
A
B
D
y
x
2
0
-2
4
0
-4
500
250
0
-250
-500
-750
-2
0
2



ADB:

Shapiro-Wilk p-values:

- A: 0.4765
- B: 0.6316
- D: < 2.2e-16

A and B appear approximately normal, but D shows a clear violation. Its p-value is near zero, indicating strong non-normality. This matches what was observed in Week 1: D has a massive spike at 0 and heavy tails. The Q-Q plot shows D's points curving away sharply from the reference line, unlike A and B which stay close to the diagonal.

Sample variances:

- A: 0.98
- B: 4.08
- D: 1253.32

The variance ratio is 1278, far above the threshold of 4. This shows extreme inequality in spread. In the residual plot, A and B have compact, symmetric scatter, but D has a much larger and uneven spread. The variability in D overwhelms the others, making any joint inference unreliable.

Both key assumptions for ANOVA — normality and equal variance — are clearly violated here due to D. Including D in any model without transformation or robust methods would lead to distorted results.

ACD:

Shapiro-Wilk p-values:

- A: 0.4765
- C: 0.5889
- D: < 2.2e-16

A and C both appear normal. D, however, again strongly violates the normality assumption. Its extremely low p-value reflects the same issue seen in earlier triplets: a sharp central spike and heavy tails. In the Q-Q plot, D curves sharply away from the theoretical line, while A and C remain nearly straight.

Sample variances:

- A: 0.98
- C: 0.96
- D: 1253.316

The max-to-min variance ratio is over 1300, confirming a massive difference in scale. A and C are tightly clustered, but D has extreme

spread. In the residual plot, residuals from A and C show uniform scatter around 0, while D's residuals are widely spread, creating a distorted shape that dominates the plot.

This triplet violates both ANOVA assumptions. D's variance is much larger than A and C, and its distribution is not Normal. Any model using these three together without transformation or robust adjustment would not be reliable.

BCD:

Shapiro-Wilk p-values:

- B: **0.6316**
- C: **0.5889**
- D: **< 2.2e-16**

B and C meet the normality assumption, with p-values comfortably above 0.05. D once again fails — its distribution is highly non-normal, just like in the other triplets. The sharp spike and long tails create a structure that departs strongly from a Normal curve.

Sample variances:

- B: **~4.08**
- C: **~0.96**
- D: **~1250+**

The variance ratio is **1309**, which reflects extreme heteroscedasticity. While B and C have moderate spreads, D is on a vastly different scale. If these three are analyzed together, the results will be dominated by D's variance, and the assumption of equal variance will be violated.

This triplet fails both assumptions. Normality is broken by D, and the huge difference in scale invalidates the equal variance condition. B and C could be compared on their own, but combining them with D without transformation would lead to misleading inferences.

Conclusion

This project showed that the patterns we saw in the histograms helped us understand the results of the later tests. Variables like A and C, which had different centers and similar spreads, were easy to tell apart and gave strong results. In contrast, variable D had very high variance and didn’t follow a normal shape, which made it harder to compare. Overall, checking shapes, spreads, and assumptions early on helped make sense of the results in t-tests, ROC curves, and ANOVA.

Appendix

HISTOGRAMS:

Histogram of A:

```
ggplot(dataset_1, aes(x=A))+geom_histogram(binwidth = 0.3, fill= "red", color =
"black")+labs(y="Frequency", x = "Value", title = "Histogram of A",)+theme_minimal()
```

Histogram of B:

```
ggplot(dataset_1, aes(x=B))+geom_histogram(binwidth = 0.4, fill= "blue", color =
"black")+labs(y="Frequency", x = "Value", title = "Histogram of B")+theme_minimal()
```

Histogram of C:

```
ggplot(dataset_1, aes(x=C))+geom_histogram(binwidth = 0.3, fill= "yellow", color =
"black")+labs(y="Frequency", x = "Value", title = "Histogram of C")+theme_minimal()
```

Histogram of D:

```
ggplot(dataset_1, aes(x=D))+geom_histogram(binwidth = 0.3, fill= "green", color =
"black")+labs(y="Frequency", x = "Value", title = "Histogram of D")+theme_minimal()
```

Overlapping Histogram of ABC:

```
ggplot(dataset_1) +geom_histogram(aes(x = A, fill = "A"), color = "black", binwidth = 0.3, alpha =
0.6) +
  geom_histogram(aes(x = B, fill = "B"), color = "black", binwidth = 0.4, alpha = 0.6) +
  geom_histogram(aes(x = C, fill = "C"), color = "black", binwidth = 0.3, alpha = 0.6) +
  labs(title = "Histogram of A, B, C", x = "Value", y = "Frequency") +
  scale_fill_manual(values = c("A" = "red", "B" = "blue", "C" = "yellow")) +   theme_minimal() +
  theme(legend.title = element_blank()) +   guides(fill = guide_legend(title = "Variables"))  # Add
custom title for the legend
```

Overlapping Histogram of ABCD:

```
ggplot(dataset_1) +
    geom_histogram(aes(x = A, fill = "A"), color = "black", binwidth = 0.3, alpha = 0.6) +
    geom_histogram(aes(x = B, fill = "B"), color = "black", binwidth = 0.4, alpha = 0.6) +
```

```
  geom_histogram(aes(x = C, fill = "C"), color = "black", binwidth = 0.3, alpha = 0.6) +
  geom_histogram(aes(x = D, fill = "D"), color = "black", binwidth = 0.5, alpha = 0.6) +
  labs(title = "Histogram of A, B, C, D", x = "Value", y = "Frequency") +
  scale_fill_manual(values = c("A" = "red", "B" = "blue", "C" = "yellow", "D" = "black")) +
theme_minimal() +
  theme(legend.title = element_blank()) +   guides(fill = guide_legend(title = "Variables"))
```

Week 2:

Transforming the data:

1) dataset_1_roc <- pivot_longer(dataset_1, cols = A:D, names_to = "Dataset", values_to = "Value") # New transformed into long format data called dataset_1_roc
2) Creating Class column to compare the Two groups:

```
roc_AB_data <- dataset_1_roc %>%
  filter(Dataset %in% c("A", "B")) %>%
  mutate(Class = ifelse(Dataset == "B", 1, 0))
roc_AC_data <- dataset_1_roc %>%
  filter(Dataset %in% c("A", "C")) %>%
  mutate(Class = ifelse(Dataset == "C", 1, 0)
roc_AD_data <- dataset_1_roc %>%
  filter(Dataset %in% c("A", "D")) %>%
  mutate(Class = ifelse(Dataset == "D", 1, 0))
roc_BC_data <- dataset_1_roc %>%
  filter(Dataset %in% c("B", "C")) %>%
  mutate(Class = ifelse(Dataset == "C", 1, 0))
roc_BD_data <- dataset_1_roc %>%
  filter(Dataset %in% c("B", "D")) %>%
  mutate(Class = ifelse(Dataset == "D", 1, 0))
roc_CD_data <- dataset_1_roc %>%
  filter(Dataset %in% c("C", "D")) %>%
  mutate(Class = ifelse(Dataset == "D", 1, 0))
```

3) Calculating ROS curve:

```
roc_AB <- roc(roc_AB_data$Class, roc_AB_data$Value)
roc_AC <- roc(roc_AC_data$Class, roc_AC_data$Value)
roc_AD <- roc(roc_AD_data$Class, roc_AD_data$Value)
```

```
roc_BC <- roc(roc_BC_data$Class, roc_BC_data$Value)
roc_BD <- roc(roc_BD_data$Class, roc_BD_data$Value)
roc_CD <- roc(roc_CD_data$Class, roc_CD_data$Value)
```

4) Graph the ROC curves:

A vs B

```
ggplot(data.frame(FPR = 1 - roc_AB$specificities, TPR =
roc_AB$sensitivities), aes(x = FPR, y = TPR))   + geom_line(linewidth
= 1.2) + geom_abline(slope = 1, intercept = 0, linetype = "dashed",
color = "gray") +  labs(title = "ROC Curve for A vs B", x = "False
Positive Rate", y = "True Positive Rate")+  theme_minimal()
```

A vs C

```
ggplot(data.frame(FPR = 1 - roc_AC$specificities, TPR =
roc_AC$sensitivities),
       aes(x = FPR, y = TPR)) +
  geom_line(linewidth = 1.2, color = "blue") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color =
"gray") +
  labs(title = "ROC Curve for A vs C",
       x = "False Positive Rate",
       y = "True Positive Rate") +
  theme_minimal()
```

A vs D

```
ggplot(data.frame(FPR = 1 - roc_AD$specificities, TPR =
roc_AD$sensitivities),
       aes(x = FPR, y = TPR)) +
  geom_line(linewidth = 1.2, color = "green") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color =
"gray") +
  labs(title = "ROC Curve for A vs D",
       x = "False Positive Rate",
       y = "True Positive Rate") +
  theme_minimal()
```

B vs C

```
ggplot(data.frame(FPR = 1 - roc_BC$specificities, TPR =
roc_BC$sensitivities),
```

```
      aes(x = FPR, y = TPR)) +
  geom_line(linewidth = 1.2, color = "purple") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color =
"gray") +
  labs(title = "ROC Curve for B vs C",
       x = "False Positive Rate",
       y = "True Positive Rate") +
  theme_minimal()
```

B vs D

```
ggplot(data.frame(FPR = 1 - roc_BD$specificities, TPR =
roc_BD$sensitivities),
       aes(x = FPR, y = TPR)) +
  geom_line(linewidth = 1.2, color = "orange") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color =
"gray") +
  labs(title = "ROC Curve for B vs D",
       x = "False Positive Rate",
       y = "True Positive Rate") +
  theme_minimal()
```

C vs D

```
ggplot(data.frame(FPR = 1 - roc_CD$specificities, TPR =
roc_CD$sensitivities),
       aes(x = FPR, y = TPR)) +
    geom_line(linewidth = 1.2, color = "red") +
    geom_abline(slope = 1, intercept = 0, linetype = "dashed", color =
"gray") +
    labs(title = "ROC Curve for C vs D",
         x = "False Positive Rate",
         y = "True Positive Rate") +
    theme_minimal()
```

All 6 PAIRS ROC:

```
plot(roc_AB, col = "black", lwd = 2, main = "ROC Curves for All
Comparisons",  legacy.axes = TRUE, xlab = "False Positive Rate",
ylab = "True Positive Rate",    bty = "n")
lines(roc_AC, col = "blue", lwd = 2)
```

```
lines(roc_AD, col = "green", lwd = 2)
lines(roc_BC, col = "purple", lwd = 2)
lines(roc_BD, col = "orange", lwd = 2)
lines(roc_CD, col = "red", lwd = 2)


legend("bottomright",
       legend = c("A vs B", "A vs C", "A vs D", "B vs C", "B vs D", "C vs
D"),
       col = c("black", "blue", "green", "purple", "orange", "red"),
       lwd = 2,
       box.lty = 0)
```

Area Under the curve(AUC):

```
 auc_AB <- auc(roc_AB)
 auc_AC <- auc(roc_AC)
 auc_AD <- auc(roc_AD)
 auc_BC <- auc(roc_BC)
 auc_BD <- auc(roc_BD)
 auc_CD <- auc(roc_CD)
 auc_AB; auc_AC; auc_AD; auc_BC; auc_BD; auc_CD
Area under the curve: 0.5163
Area under the curve: 0.7571
Area under the curve: 0.698
Area under the curve: 0.65
Area under the curve: 0.6297
Area under the curve: 0.5171
```

Week 3:

Two Sample T-test

```
pairs <- list(
  c("A", "B"),
  c("A", "C"),
```

```
  c("A", "D"),
  c("B", "C"),
  c("B", "D"),
  c("C", "D"))

for (pair in pairs) {col1 <- pair[1], col2 <- pair[2]

    data_pair <- data.frame(Value = c(dataset_1[[col1]], dataset_1[[col2]]),
    Group = factor(rep(c(col1, col2), each = nrow(dataset_1))))

   t_test_result <- t.test(Value ~ Group, data = data_pair)
  cat("\n==============================\n")
  cat("T-test result for", col1, "vs", col2, ":\n")
  print(t_test_result)}
```

Graphs:

A vs B:

```
data_pair <- data.frame(
  Value = c(dataset_1$A, dataset_1$B),
  Group = factor(rep(c("A", "B"), each = nrow(dataset_1)))
)

ggplot(data_pair, aes(x = Group, y = Value, fill = Group)) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Comparison: A vs B", y = "Value", x = "") +
  theme_minimal()
```

T-test result for A vs B :

```
        Welch Two Sample t-test

data:  Value by Group
t = -0.96711, df = 1454.4, p-value = 0.3336
alternative hypothesis: true difference in means between group A and
group B is not equal to 0
95 percent confidence interval:
```

-0.20835512 0.07074984
sample estimates:
mean in group A mean in group B
     0.01612787   0.08493051

A vs C:

```
data_pair <- data.frame(
  Value = c(dataset_1$A, dataset_1$C),
  Group = factor(rep(c("A", "C"), each = nrow(dataset_1))))
ggplot(data_pair, aes(x = Group, y = Value, fill = Group)) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Comparison: A vs C", y = "Value", x = "") +
  theme_minimal()
```

T-test result for A vs C :

```
        Welch Two Sample t-test

data:  Value by Group
t = -21.877, df = 1997.6, p-value < 2.2e-16
alternative hypothesis: true difference in means between group A and
group C is not equal to 0
95 percent confidence interval:
 -1.0501537 -0.8773655
sample estimates:
mean in group A mean in group C
     0.01612787       0.97988747
```

A vs D:

```
data_pair <- data.frame(
  Value = c(dataset_1$A, dataset_1$D),
  Group = factor(rep(c("A", "D"), each = nrow(dataset_1))))
ggplot(data_pair, aes(x = Group, y = Value, fill = Group)) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Comparison: A vs D", y = "Value", x = "") +
  theme_minimal()
```

T-test result for A vs D :

```
        Welch Two Sample t-test

data:  Value by Group
t = -1.5223, df = 1000.6, p-value = 0.1283
alternative hypothesis: true difference in means between group A and
group D is not equal to 0
95 percent confidence interval:
 -3.9026172  0.4928435
sample estimates:
mean in group A mean in group D
     0.01612787      1.72101475
```

B vs C:

```
data_pair <- data.frame(
  Value = c(dataset_1$B, dataset_1$C),
  Group = factor(rep(c("B", "C"), each = nrow(dataset_1)))
)

ggplot(data_pair, aes(x = Group, y = Value, fill = Group)) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Comparison: B vs C", y = "Value", x = "") +
  theme_minimal()
```

T-test result for B vs C :

```
        Welch Two Sample t-test

data:  Value by Group
t = -12.613, df = 1443.5, p-value < 2.2e-16
alternative hypothesis: true difference in means between group B and
group C is not equal to 0
95 percent confidence interval:
 -1.0341476 -0.7557663
sample estimates:
mean in group B mean in group C
```

```
    0.08493051      0.97988747
```

B vs D:

```
data_pair <- data.frame(
  Value = c(dataset_1$B, dataset_1$D),
  Group = factor(rep(c("B", "D"), each = nrow(dataset_1))))

ggplot(data_pair, aes(x = Group, y = Value, fill = Group)) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Comparison: B vs D", y = "Value", x = "") +
  theme_minimal()
```

T-test result for B vs D :

```
        Welch Two Sample t-test

data:  Value by Group
t = -1.459, df = 1005.5, p-value = 0.1449
alternative hypothesis: true difference in means between group B and
group D is not equal to 0
95 percent confidence interval:
 -3.8365107  0.5643422
sample estimates:
mean in group B mean in group D
    0.08493051      1.72101475
```

C vs D:

```
data_pair <- data.frame(
  Value = c(dataset_1$C, dataset_1$D),
  Group = factor(rep(c("C", "D"), each = nrow(dataset_1))))

ggplot(data_pair, aes(x = Group, y = Value, fill = Group)) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Comparison: C vs D", y = "Value", x = "") +
  theme_minimal()
```

T-test result for C vs D :

```
        Welch Two Sample t-test

data:  Value by Group
t = -0.66175, df = 1000.5, p-value = 0.5083
alternative hypothesis: true difference in means between group C and
group D is not equal to 0
95 percent confidence interval:
 -2.938835  1.456580
sample estimates:
mean in group C mean in group D
      0.9798875       1.7210147
```

Week 4:

```
long_data <- pivot_longer(dataset_1, cols = c("A", "B", "C", "D"),
                          names_to = "Group", values_to = "Value")
Model 1:
abc_data <- subset(long_data, Group %in% c("A", "B", "C"))
abc_model <- aov(Value ~ Group, data = abc_data)
summary(abc_model)

TukeyHSD(abc_model)
boxplot(Value ~ Group, data = abc_data,
        main = "ANOVA: A vs B vs C",
        xlab = "Group", ylab = "Value",
        col = c("red", "blue", "green"))
Model 2:
bcd_data <- subset(long_data, Group %in% c("B", "C", "D"))
bcd_model <- aov(Value ~ Group, data = bcd_data)
summary(bcd_model)
boxplot(Value ~ Group, data = bcd_data,
        main = "ANOVA: B vs C vs D",
        xlab = "Group", ylab = "Value",
        col = c("blue", "green", "black"))
```

```
Model 1:
  Df Sum Sq Mean Sq F value Pr(>F)
Group         2    578  289.09   144.1 <2e-16 ***
Residuals   2997   6012    2.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


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

Fit: aov(formula = Value ~ Group, data = abc_data)

$Group
         diff         lwr       upr     p adj
B-A 0.06880264 -0.07972717 0.2173325 0.5227061
C-A 0.96375961  0.81522979 1.1122894 0.0000000
C-B 0.89495697  0.74642715 1.0434868 0.0000000


Model 2:
Df  Sum Sq Mean Sq F value Pr(>F)
Group         2    1342   671.2     1.6  0.202
Residuals   2997 1257093   419.5
```

Week 5:

ABC

```
abc_data <- subset(long_data, Group %in% c("A", "B", "C"))

shapiro.test(abc_data$Value[abc_data$Group == "A"])
```

```
Shapiro-Wilk normality test

data:  abc_data$Value[abc_data$Group == "A"]
W = 0.99838, p-value = 0.4765

shapiro.test(abc_data$Value[abc_data$Group == "B"])
Shapiro-Wilk normality test

data:  abc_data$Value[abc_data$Group == "B"]
W = 0.99862, p-value = 0.6316

shapiro.test(abc_data$Value[abc_data$Group == "C"])
Shapiro-Wilk normality test

data:  abc_data$Value[abc_data$Group == "C"]
W = 0.99855, p-value = 0.5889

Variance:
var_A <- var(abc_data$Value[abc_data$Group == "A"])
[1] 0.9834589
var_B <- var(abc_data$Value[abc_data$Group == "B"])
[1] 4.077768
var_C <- var(abc_data$Value[abc_data$Group == "C"])
[1] 0.9571833
max(var_A, var_B, var_C) / min(var_A, var_B, var_C)
[1] 4.260174
QQ plot:
ggplot(abc_data, aes(sample = Value)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ Group, scales = "free") +
  theme_minimal() +
  labs(title = "Q-Q Plots for A, B, C", x = "Theoretical Quantiles", y =
"Sample Quantiles")
```

Residual vs Group Plot:

```
abc_model <- aov(Value ~ Group, data = abc_data)
abc_data$residuals <- residuals(abc_model)
ggplot(abc_data, aes(x = Group, y = residuals)) +
  geom_jitter(width = 0.2, alpha = 0.5, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Group (A, B, C)",
       x = "Group (X)", y = "Residuals (e_i)") +
  theme_minimal()
```

ABD

```
abd_data <- subset(long_data, Group %in% c("A", "B", "D"))
```

- shapiro.test(abd_data$Value[abd_data$Group == "A"])

```
Shapiro-Wilk normality test
W = 0.99838, p-value = 0.4765
```

- shapiro.test(abd_data$Value[abd_data$Group == "B"])

```
Shapiro-Wilk normality test
W = 0.99862, p-value = 0.6316
```

- shapiro.test(abd_data$Value[abd_data$Group == "D"])

```
Shapiro-Wilk normality test
W = 0.13974, p-value < 2.2e-16
```

```
var_A <- var(abd_data$Value[abd_data$Group == "A"])
var_B <- var(abd_data$Value[abd_data$Group == "B"])
var_D <- var(abd_data$Value[abd_data$Group == "D"])
max(var_A, var_B, var_D) / min(var_A, var_B, var_D)
[1] 1274.396
```

Residual Plot:

```
abd_model <- aov(Value ~ Group, data = abd_data)
abd_data$residuals <- residuals(abd_model)

ggplot(abd_data, aes(x = Group, y = residuals)) +
  geom_jitter(width = 0.2, alpha = 0.5, color = "darkred") +
```

```
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Group (A, B, D)",
       x = "Group", y = "Residuals") +
  theme_minimal()
QQ plot:
ggplot(abd_data, aes(sample = Value)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ Group, scales = "free") +
  theme_minimal() +
  labs(title = "Q-Q Plots for A, B, D")
```

ACD

```
acd_data <- subset(long_data, Group %in% c("A", "C", "D"))

shapiro.test(acd_data$Value[acd_data$Group == "A"])
W = 0.99838, p-value = 0.4765
shapiro.test(acd_data$Value[acd_data$Group == "C"])
W = 0.99855, p-value = 0.5889
shapiro.test(acd_data$Value[acd_data$Group == "D"])
W = 0.13974, p-value < 2.2e-16

var_A <- var(acd_data$Value[acd_data$Group == "A"])
var_C <- var(acd_data$Value[acd_data$Group == "C"])
var_D <- var(acd_data$Value[acd_data$Group == "D"])

max(var_A, var_C, var_D) / min(var_A, var_C, var_D)
[1] 1309.379

acd_model <- aov(Value ~ Group, data = acd_data)
acd_data$residuals <- residuals(acd_model)

ggplot(acd_data, aes(x = Group, y = residuals)) +
  geom_jitter(width = 0.2, alpha = 0.5, color = "purple") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Group (A, C, D)",
       x = "Group", y = "Residuals") +
```

```
  theme_minimal()
```

BCD

```
bcd_data <- subset(long_data, Group %in% c("B", "C", "D"))

shapiro.test(bcd_data$Value[bcd_data$Group == "B"])
W = 0.99862, p-value = 0.6316
shapiro.test(bcd_data$Value[bcd_data$Group == "C"])
W = 0.99855, p-value = 0.5889
shapiro.test(bcd_data$Value[bcd_data$Group == "D"])
W = 0.13974, p-value < 2.2e-16

var_B <- var(bcd_data$Value[bcd_data$Group == "B"])
var_C <- var(bcd_data$Value[bcd_data$Group == "C"])
var_D <- var(bcd_data$Value[bcd_data$Group == "D"])


max(var_B, var_C, var_D) / min(var_B, var_C, var_D) = 1309.379
```