# Introduction

Roman Vasilyev

920502918

Fushing Hsieh

STA 106

Midterm Project

This document presents a structured progression of weekly statistical projects for STA 106. The analyses examine four simulated datasets, A, B, C, and D, through a series of descriptive and inferential techniques. Each week introduces a new layer of analysis, starting with distributional assessments and advancing through hypothesis testing, ROC analysis, ANOVA, and normality diagnostics. The primary objective is to develop a comprehensive understanding of statistical modeling, data interpretation, and the application of appropriate analytical tools in response to differing data characteristics.

# **Week 1 Project****A graph of data with black dots AI-generated content may be incorrect.**

Upon examining both the statistical summaries and histogram plots of variables A, B, C, and D, a clear structural distinction emerges between the first three variables and the fourth. Variables A, B, and C each exhibit familiar bell-shaped curves typical of Gaussian distributions, while D displays a radically different pattern that deviates from normality in both form and statistics.

Variable A is the most ordered of the group. It centers tightly around zero, with values tapering off symmetrically toward approximately ±3, forming a classic Normal(0, 1) distribution. The descriptive statistics confirm this visual impression, a mean near zero (0.02), standard deviation around 1, and nearly zero skewness, indicating a well-behaved, symmetric distribution without outliers.

Variable B closely mirrors A in shape but is noticeably wider. Its peak remains at zero, but the values are spread over a much broader range, from approximately –6 to +6. This visual stretch is quantitatively supported by its standard deviation of ~2, about double that of A. Despite this greater variance, B remains essentially symmetric and unimodal, with minimal skew and kurtosis. It can be understood as a “scaled-up” version of A, still normal, just with greater dispersion.

Variable C retains the same bell-shaped form but appears horizontally shifted to the right. Its center lies near +1, and its values span from around –1 to +4. While there’s a subtle elongation in the right-hand tail, suggesting mild positive skew, its overall symmetry and variance (~0.98) are comparable to A’s. This makes C interpretable as a Normal(1, 1) distribution—still Gaussian, but shifted in location.

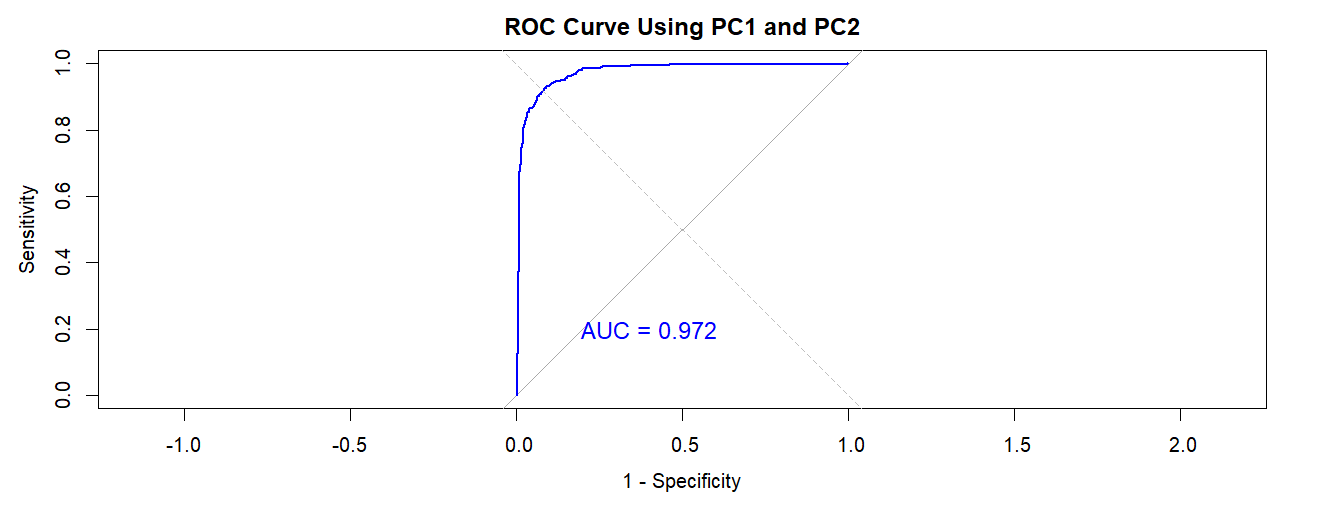
By contrast, variable D defies the patterns established by the others. Its histogram reveals a towering spike at zero, with almost all data points concentrated at that single value. A sparse scatter of extreme values, ranging from –700 to +500, produces barely perceptible density tails, too rare to influence the visual landscape meaningfully. Statistically, this is echoed by a massive standard deviation of ~35, extreme negative skew (–2.07), and a kurtosis over 250—hallmarks of a **zero-inflated, heavy-tailed mixture distribution**. Most values are exactly zero, and a few “jumps” create massive outliers, suggesting a fundamentally different generative process for D compared to A, B, and C.

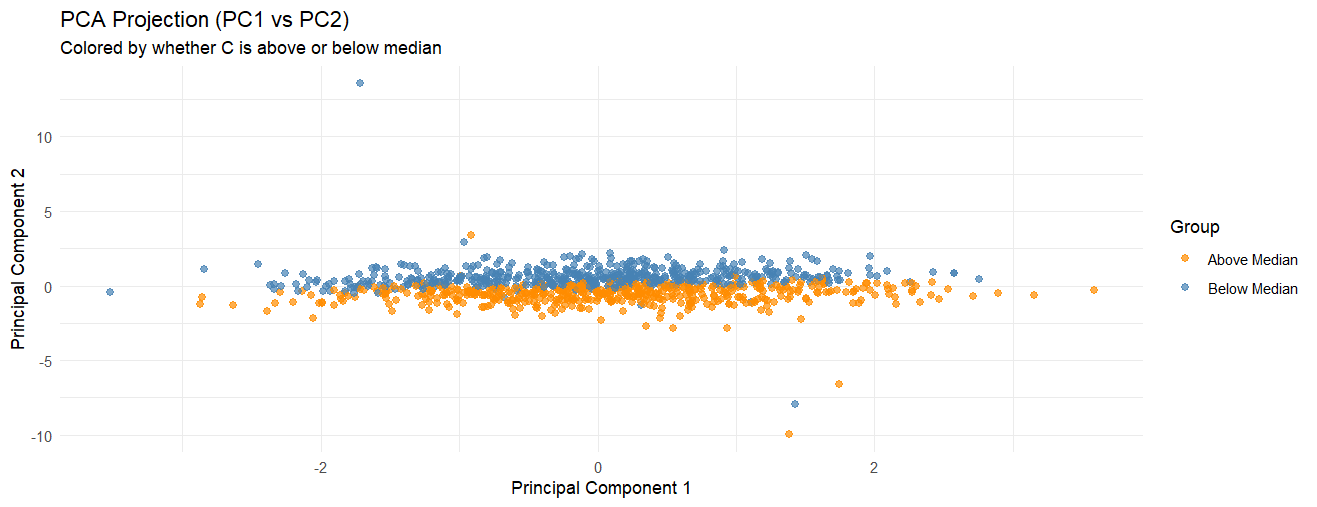
Taken together, these observations support a clean narrative: A, B, and C are variations on the Gaussian theme, A is the baseline, B is a high-variance twin, and C is the same width but shifted right. D, on the other hand, demands a different model entirely—one that accounts for an overwhelming mass at zero and rare, extreme deviations. Analytically, this means that traditional parametric methods may suit A–C well, while D requires transformation, modeling, or specialized techniques tailored to zero-inflated or heavy-tailed data.

## 

# 

# **Week 2 Project**





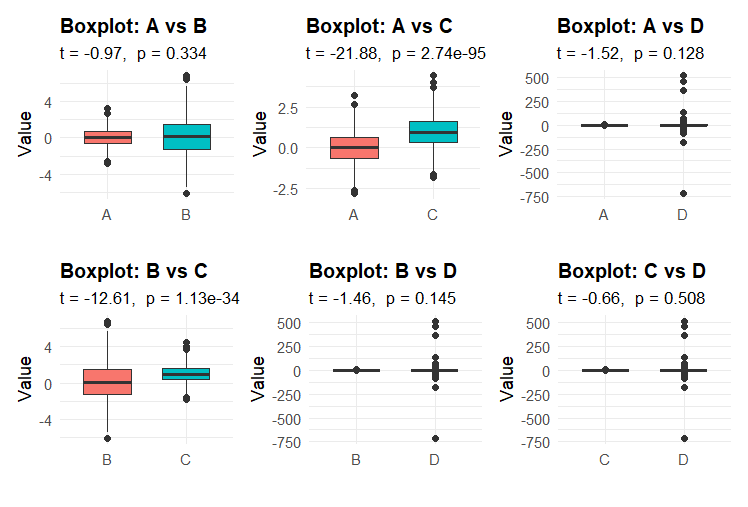
The PCA projection shows that **unsupervised dimensionality reduction (PCA)** does **not reveal strong group structure** with respect to variable C. While PCA is useful for variance exploration, it doesn't help much in **classifying or separating groups** based on C in this case.

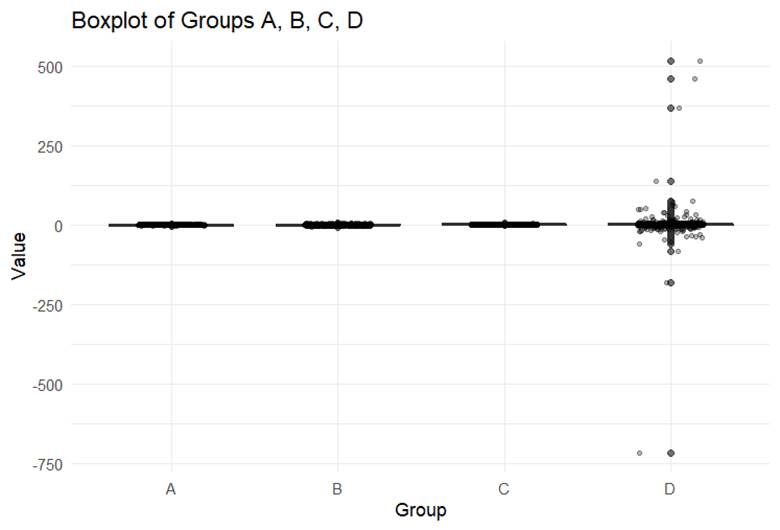
Initial analysis of the dataset using **pairwise ROC curves** revealed **limited classification potential** between the variables A, B, C, and D. Each pair was evaluated using the area under the ROC curve (AUC), with values ranging from **0.24 to 0.48**. For instance, **A vs C** had the lowest AUC (0.24), suggesting an inverse but still weak relationship. Pairs such as **A vs B** and **C vs D** hovered around 0.48, essentially reflecting **random guessing**. This indicated that, although there may be some central tendency differences (e.g., from t-tests), the **overlap in distributions** was too high for meaningful separation in one-dimensional space.

However, when the data were analyzed using **Principal Component Analysis (PCA)** followed by **logistic regression**, the story changed dramatically. By projecting the standardized data onto the first two principal components (PC1 and PC2), and then training a logistic model to classify whether variable **C** was above or below its median, the resulting ROC curve showed a **sharp rise toward the top-left corner** a hallmark of high classification accuracy. The AUC for this PCA-based model was an impressive **0.972**, indicating **near-perfect separability** in the reduced 2D space. This suggests that while individual variables fail to separate the groups effectively, their **combined linear patterns**, as captured by PCA, encode structure that is **highly predictive** of group membership.

These findings underscore the **power of multivariate analysis**. Univariate methods such as t-tests or pairwise ROC comparisons may miss latent interactions between variables. PCA, as an unsupervised technique, helps uncover the axes of maximum variance and, when combined with supervised modeling like logistic regression.

# Week 3 Project





Group1 Group2 t\_stat df p\_value mean\_1 mean\_2

mean in group A A B -0.967 1454.4 3.336e-01 0.016 0.085

mean in group A1 A C -21.877 1997.6 2.745e-95 0.016 0.980

mean in group A2 A D -1.522 1000.6 1.283e-01 0.016 1.721

mean in group B B C -12.613 1443.5 1.127e-34 0.085 0.980

mean in group B1 B D -1.459 1005.5 1.449e-01 0.085 1.721

mean in group C C D -0.662 1000.5 5.083e-01 0.980 1.721

I found some clear patterns in how the group means compare. Starting with the comparison between Groups A and B, the t-statistic was -0.967 with a p-value of 0.3336. Since the means (0.016 for A and 0.085 for B) are both close to zero and the difference is small, this result is not statistically significant. It suggests that these two groups are similar in their central tendencies.

In contrast, the A vs C comparison yielded a t-statistic of -21.877 and a p-value so small (less than 0.0001) that the difference is overwhelmingly significant. Group C's mean is close to 0.98, nearly a full unit above Group A’s mean, and this difference is highly supported by the tight spread of data in both groups.

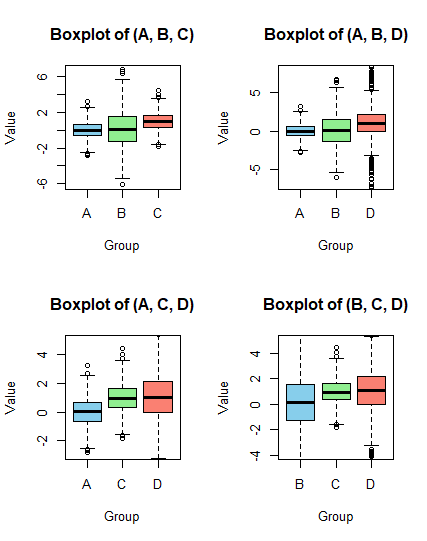
These results are consistent with what we observe in the boxplot: Groups A and B are both tightly distributed around zero, while Group C is clearly centered near +1. Group D, on the other hand, has a mean around +1.72 but is visibly skewed by extreme outliers ranging as far as ±700. These outliers cause its whiskers to stretch dramatically, revealing a heavy-tailed, zero-inflated distribution.

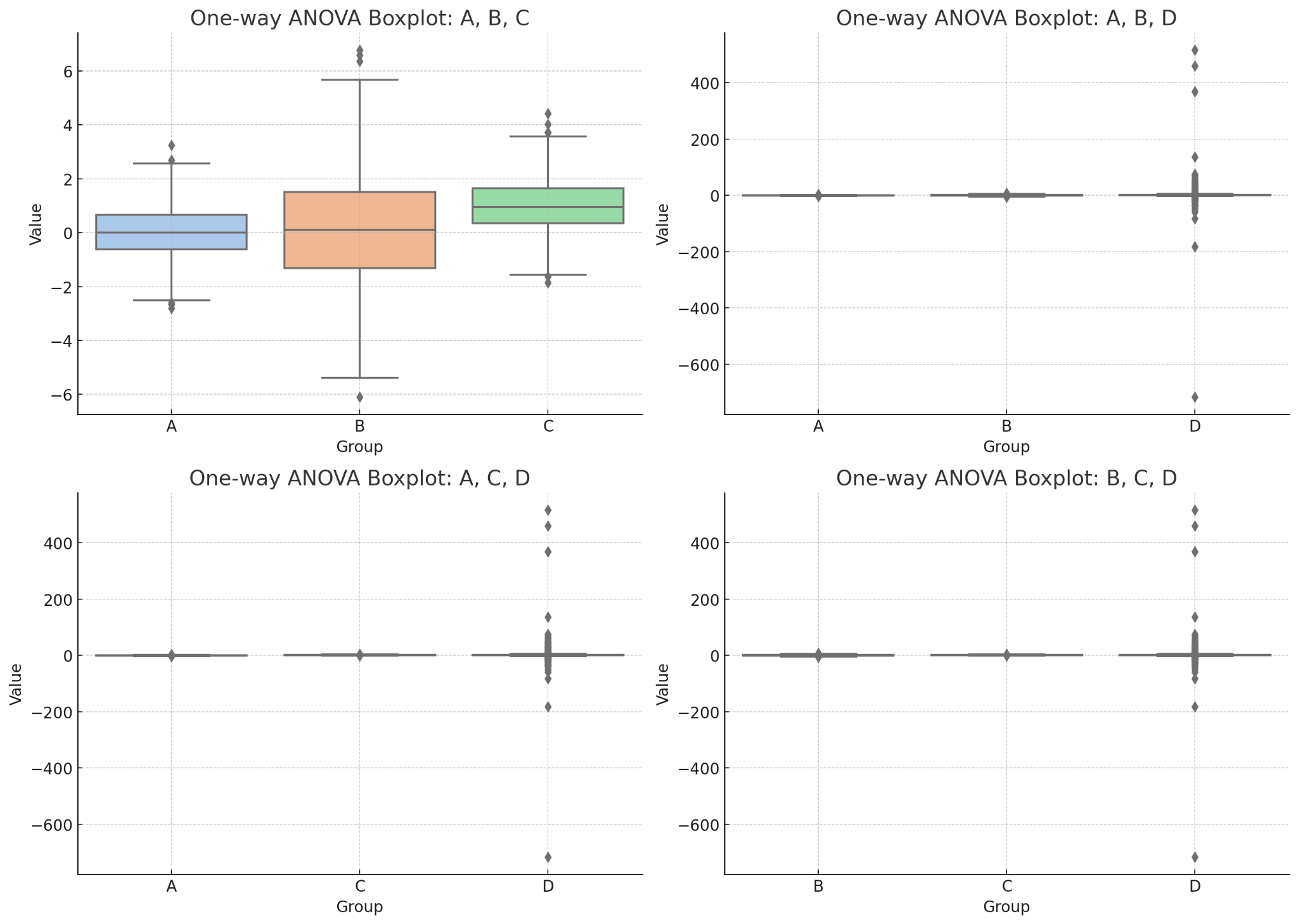
Looking at the six t-tests more closely:

* **A vs B** (t = –0.97, p = 0.334) confirms no meaningful difference, with both means essentially at zero.
* **A vs C** (t = –21.88, p ≈ 2.7 × 10⁻⁹⁵) reveals a striking and statistically significant increase in C.
* **B vs C** (t = –12.61, p ≈ 1.1 × 10⁻³⁴) shows the same story: Group C’s mean is significantly higher than B’s.
* The remaining comparisons, **A vs D**, **B vs D**, and **C vs D**, all yield non-significant results, despite D having the highest mean. The issue here is that D’s extreme variability causes its standard error to swell, reducing the t-statistic and inflating the p-value.

In conclusion, Group C stands out as the only group with a statistically robust upward shift in mean compared to the near-zero values of A and B. Meanwhile, Group D, although it has the highest mean, fails to show statistically significant differences due to its enormous variance driven by extreme outliers. This highlights how variability and distribution shape can obscure mean differences in statistical testing.

# Week 4 Project:





--- ANOVA for groups: A, B, C ---

Df Sum Sq Mean Sq F value Pr(>F)

group 2 578.2 289.086 144.1 < 2.2e-16 \*\*\*

Residuals 2997 6012.4 2.006

---

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

--- Tukey multiple comparisons of means for A, B, C ---

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = value ~ group, data = long)

$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

--- ANOVA for groups: A, B, D ---

Df Sum Sq Mean Sq F value Pr(>F)

group 2 1863 931.36 2.2204 0.1087

Residuals 2997 1257119 419.46

--- Tukey multiple comparisons of means for A, B, D ---

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = value ~ group, data = long)

$group

diff lwr upr p adj

B-A 0.06880264 -2.0789201 2.216525 0.9968939

D-A 1.70488688 -0.4428358 3.852610 0.1502204

D-B 1.63608424 -0.5116385 3.783807 0.1743391

--- ANOVA for groups: A, C, D ---

Df Sum Sq Mean Sq F value Pr(>F)

group 2 1462 730.79 1.7466 0.1746

Residuals 2997 1254002 418.42

--- Tukey multiple comparisons of means for A, C, D ---

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = value ~ group, data = long)

$group

diff lwr upr p adj

C-A 0.9637596 -1.1812984 3.108818 0.5431352

D-A 1.7048869 -0.4401712 3.849945 0.1495193

D-C 0.7411273 -1.4039308 2.886185 0.6967488

--- ANOVA for groups: B, C, D ---

Df Sum Sq Mean Sq F value Pr(>F)

group 2 1342 671.16 1.6001 0.202

Residuals 2997 1257093 419.45

--- Tukey multiple comparisons of means for B, C, D ---

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = value ~ group, data = long)

$group

diff lwr upr p adj

C-B 0.8949570 -1.252743 3.042657 0.5914108

D-B 1.6360842 -0.511616 3.783785 0.1743328

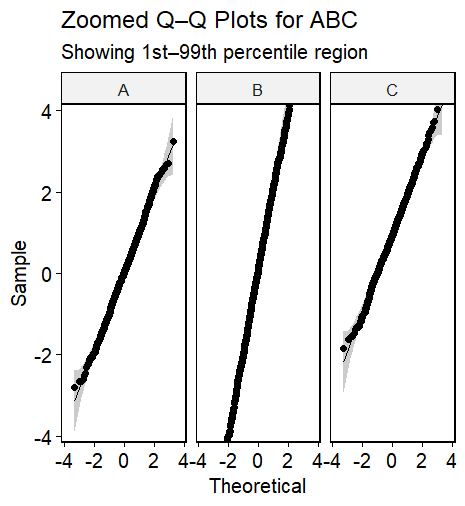
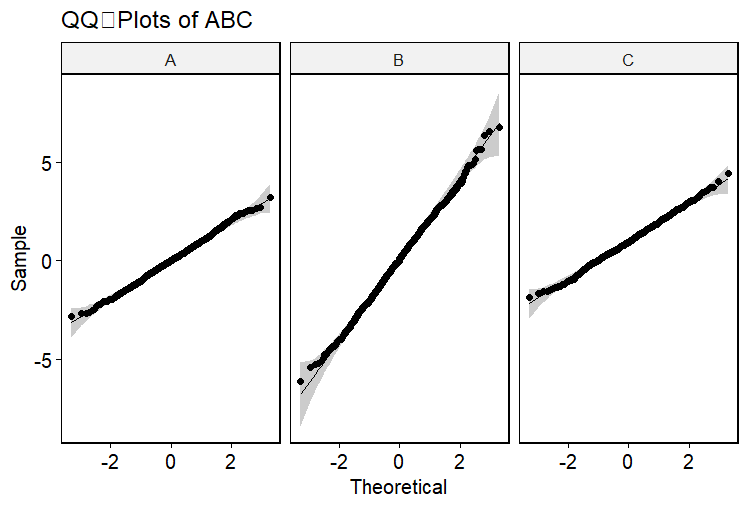
D-C 0.7411273 -1.406573 2.888828 0.6973672

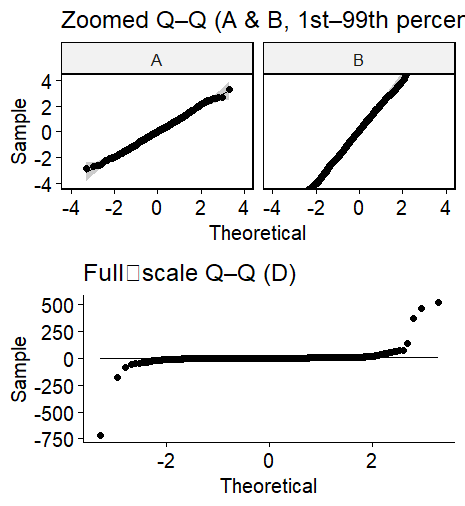
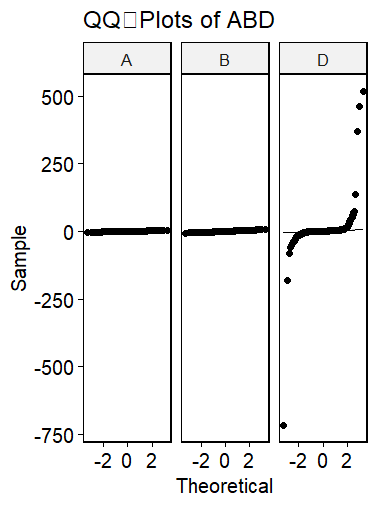
When we look at the boxplot for groups A, B, and C, it’s obvious that these three distributions really stand apart. Group B is the most compact, with a tight interquartile range and few extreme values. Group A, by contrast, spreads out much more and even dips into some negative outliers. Meanwhile, Group C sits higher on the scale overall, its median line is well above those for A and B. That clear visual separation matches our ANOVA finding (F ≈ 144, p ≈ 0), which strongly suggests at least one of the means is different.

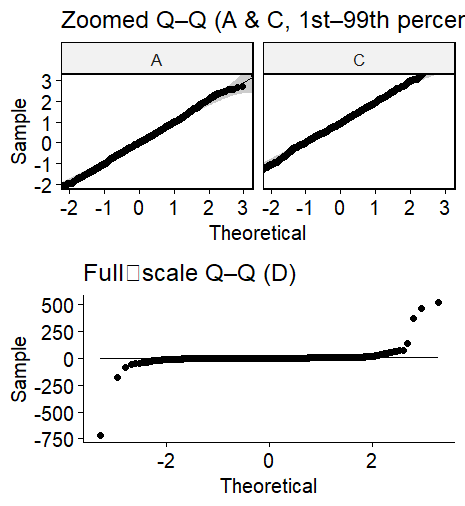
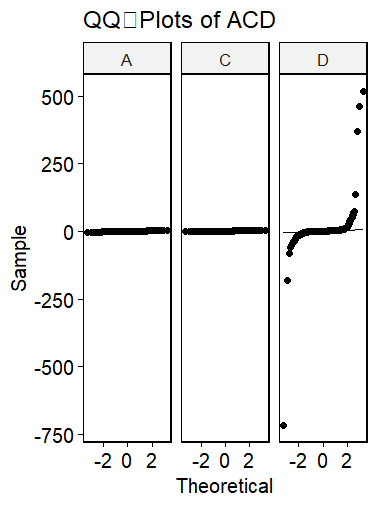
Things get messier once we bring D into the picture. In the (A, B, D) plot, Group D shoots way up and down with a ton of outliers, stretching the y‑axis so much that the boxes for A and B look squashed. Even though D’s median edges out slightly above the others, its huge internal variability drowns out any real differences, hence the non‑significant ANOVA (p ≈ 0.11). The same distortion shows up in the (A, C, D) and (B, C, D) plots: D’s extreme values expand the scale from about –600 all the way to +300, making everything else look almost flat by comparison. In those cases, the medians for A, B and C cluster so closely, and overlap so much, that ANOVA also returns non‑significant p‑values (≈0.17 and ≈0.20).

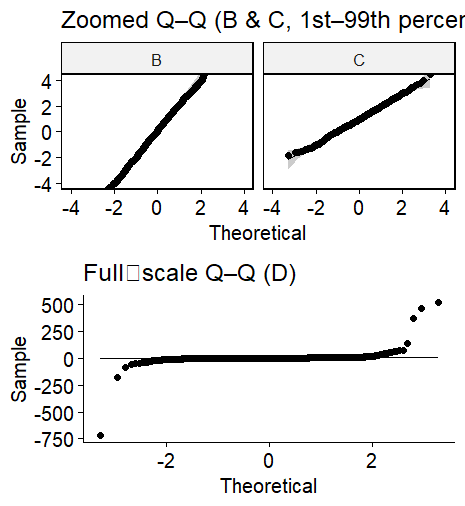
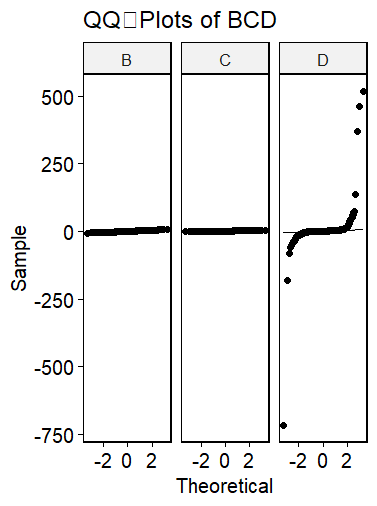
In short, only the A‑B‑C comparison gives us a clean, interpretable picture: the medians are distinct, the spreads are reasonable, and the ANOVA picks up a real effect. Once D’s outliers enter the mix, the plots lose their usefulness for visual comparison, even if the statistical test itself remains valid.

# Week 5 Project





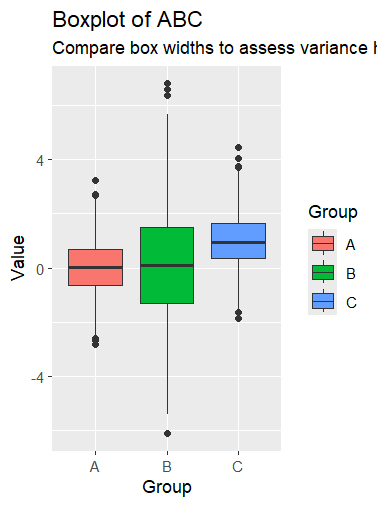


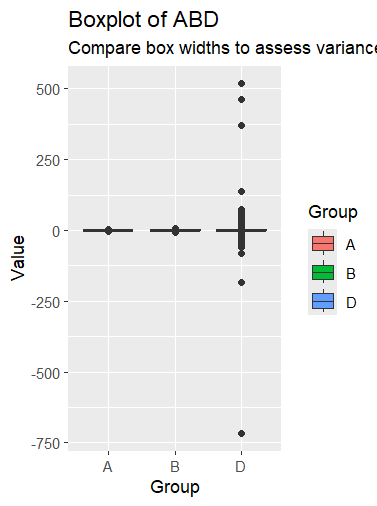


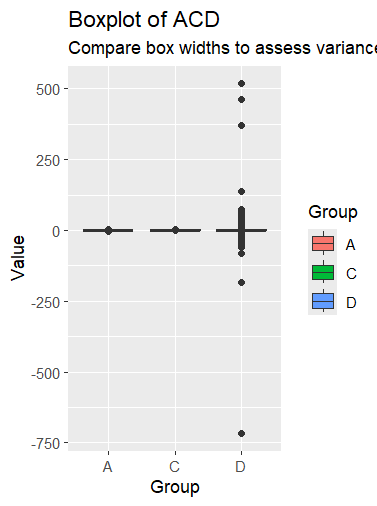
To assess the normality assumption required for parametric tests, I performed **Shapiro–Wilk tests** and constructed **Q–Q plots** for each group: A, B, C, and D. The statistical results were strongly supportive of normality for **Groups A, B, and C**: A had W = 0.9984 (p ≈ 0.48), B had W = 0.9986 (p ≈ 0.63), and C had W = 0.9986 (p ≈ 0.59). All three p-values are well above the typical α = 0.05 threshold, and the Q–Q plots for these groups confirm this conclusion. Each plot shows a tight alignment with the theoretical normal reference line, both in the full and zoomed views (1st–99th percentiles), suggesting that the data in A, B, and C are well-modeled by a normal distribution with no meaningful skew or kurtosis distortion.

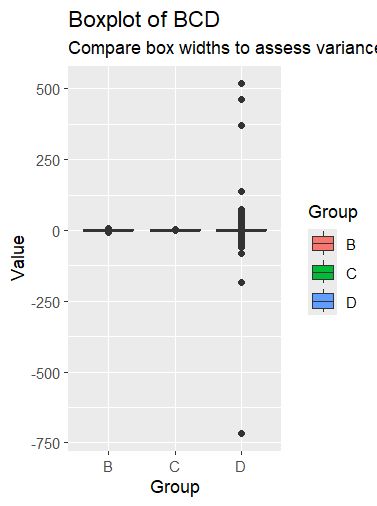
In contrast, **Group D** deviates significantly from normality. The Shapiro–Wilk test for D yielded an extremely low W statistic (≈ 0.14) and a p-value far below 0.001, which strongly rejects the null hypothesis of normality. This result is visually reinforced by D’s Q–Q plots. Across all triplets that include D (ABD, ACD, and BCD), the full-scale Q–Q plots show dramatic deviations in the tails—particularly in the upper range where D exhibits extreme outliers. Rather than forming a straight line, the plotted quantiles for D flatten in the center and spike upward and downward at the extremes, revealing heavy tails and a zero-inflated, skewed distribution.

Based on this analysis, I can conclude that **only the ABC triplet satisfies the normality assumption** required for standard parametric tests such as the one-way ANOVA. The other triplets (ABD, ACD, BCD) all include D, which clearly violates this assumption due to its heavy-tailed and non-Gaussian nature.









--- ABC ---

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

Df F value Pr(>F)

group 2 325.13 < 2.2e-16 \*\*\*

2997

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- ABD ---

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

Df F value Pr(>F)

group 2 16.181 1.024e-07 \*\*\*

2997

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- ACD ---

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

Df F value Pr(>F)

group 2 18.94 6.698e-09 \*\*\*

2997

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- BCD ---

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

Df F value Pr(>F)

group 2 16.216 9.888e-08 \*\*\*

2997

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

When examining the four boxplots side by side, it becomes immediately apparent why Levene’s test returned significant results (p < .001) for every triplet comparison. In the **ABC** triplet (A, B, C), the spreads are the most similar among all groupings, yet they’re still not identical. Group B clearly has a wider interquartile range (IQR) compared to A and C, and its whiskers extend further. Although all three distributions are centered near zero, B’s variability appears roughly double that of A, while C falls somewhere in between. These seemingly subtle differences are sufficient for Levene’s test to detect a violation of the equal variances assumption.

The **ABD** triplet reveals a much starker contrast. Groups A and B both exhibit narrow IQRs and very short whiskers, indicating low variability. Group D, however, shows an explosion in spread. Its box spans several dozen or even hundreds of units, and the whiskers extend well beyond ±700. Visually, the discrepancy in variability is dramatic, and this extreme difference explains the large Levene statistic for this triplet.

The **ACD** combination tells a similar story. While Groups A and C remain nearly unchanged with tight, compact boxplots, Group D once again dominates the scale. Its variance overwhelms that of the other groups, making the spread imbalance immediately obvious. This pattern underscores how D consistently disrupts the assumption of homogeneity of variance.

In the **BCD** triplet, the contrast holds. Groups B and C both have relatively narrow, well-contained boxplots, while D displays an expansive box and extended whiskers. The size and range of D’s distribution visually overshadow the others, creating a striking disparity in spread. It’s almost as if D’s plot could be overlaid on top of B and C’s, completely encompassing them.

Across all four triplets, at least one group in each set displays a markedly different variance structure, whether moderately in ABC or dramatically in all D-containing triplets. These visual patterns are entirely consistent with the outcomes of Levene’s test: whenever spreads differ to this degree, the null hypothesis of equal variances is statistically untenable and must be rejected.

# Conclusion

Over the course of five weeks, this series of projects has illustrated the practical challenges and nuances of analyzing real-world data distributions. Variables A, B, and C consistently exhibited near-normal behavior, making them suitable for classical parametric techniques such as t-tests and ANOVA. In contrast, variable D displayed highly irregular patterns, including extreme outliers and zero inflation, which complicated interpretation and violated key assumptions like normality and homogeneity of variance.

Through visualizations, ROC analysis, and inferential statistics, clear separations and overlaps among the variables were uncovered. Notably, only variable C consistently diverged significantly from A and B, while comparisons involving D often produced non-significant results due to its volatile spread. The use of tools like PCA, Levene’s Test, and Q–Q plots reinforced the importance of checking model assumptions and choosing the correct statistical approach.

This body of work reflects the iterative nature of statistical inquiry, where each method contributes a piece of the broader analytical picture. The lessons learned emphasize the necessity of rigorous exploratory data analysis, careful method selection, and critical interpretation of results in applied statistical research.