phuse connect
connect·share·advance
The Clinical Data Science Conference
US 2025
16–19 March
Renaissance Orlando at SeaWorld Florida

Paper ET11

Advanced Clinical Data Visualisation Using Violin Plots in R

Mrityunjay Kumar, Ephicacy Lifescience Analytics, Bengaluru, India
Shashikant Kumar, Ephicacy Lifescience Analytics, Bengaluru, India

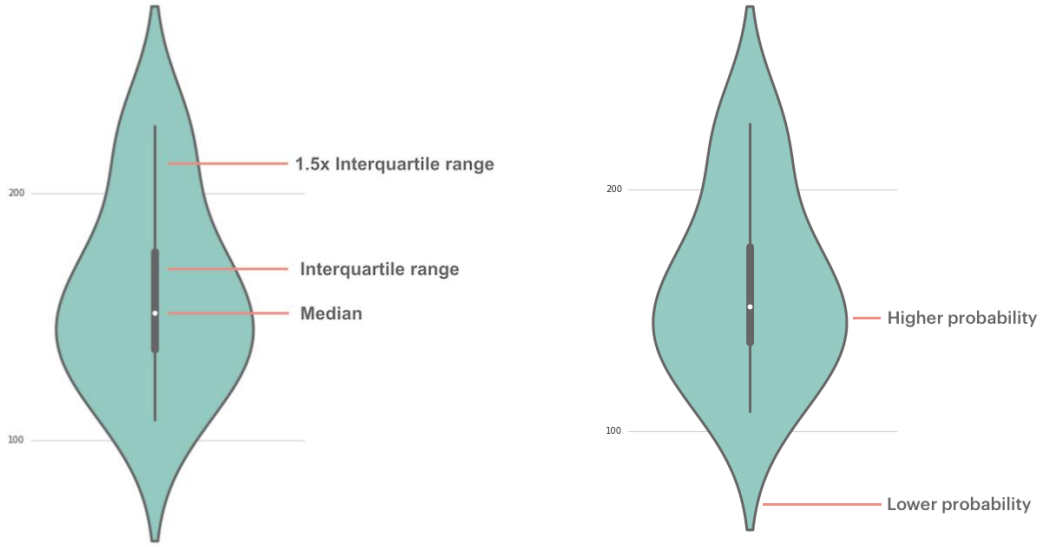
Abstract

Effective visualisation of clinical data is crucial for insightful analysis and decision-making. The violin plot, a combination of box plot and density plot, offers comprehensive visualisation by displaying data distribution and variability in a single, cohesive format. This paper presents the application of violin plots using R, showcasing how to visualise clinical trial data, emphasizing its utility in highlighting data nuances, often overlooked by traditional methods. Attendees will learn the process of creating and customizing violin plots with *ggplot2*, using real-world examples to highlight how violin plots can effectively communicate complex data patterns, such as patient response variability, subgroup comparisons, and treatment effects. By leveraging violin plots, programmers or statisticians can reveal data nuances, such as skewness, multimodal distributions and outliers leading to more informed and impactful data-driven decisions. This paper aims to provide participants with required details for creation of violin plots, thus enhancing the quality of their visualisations.

Introduction

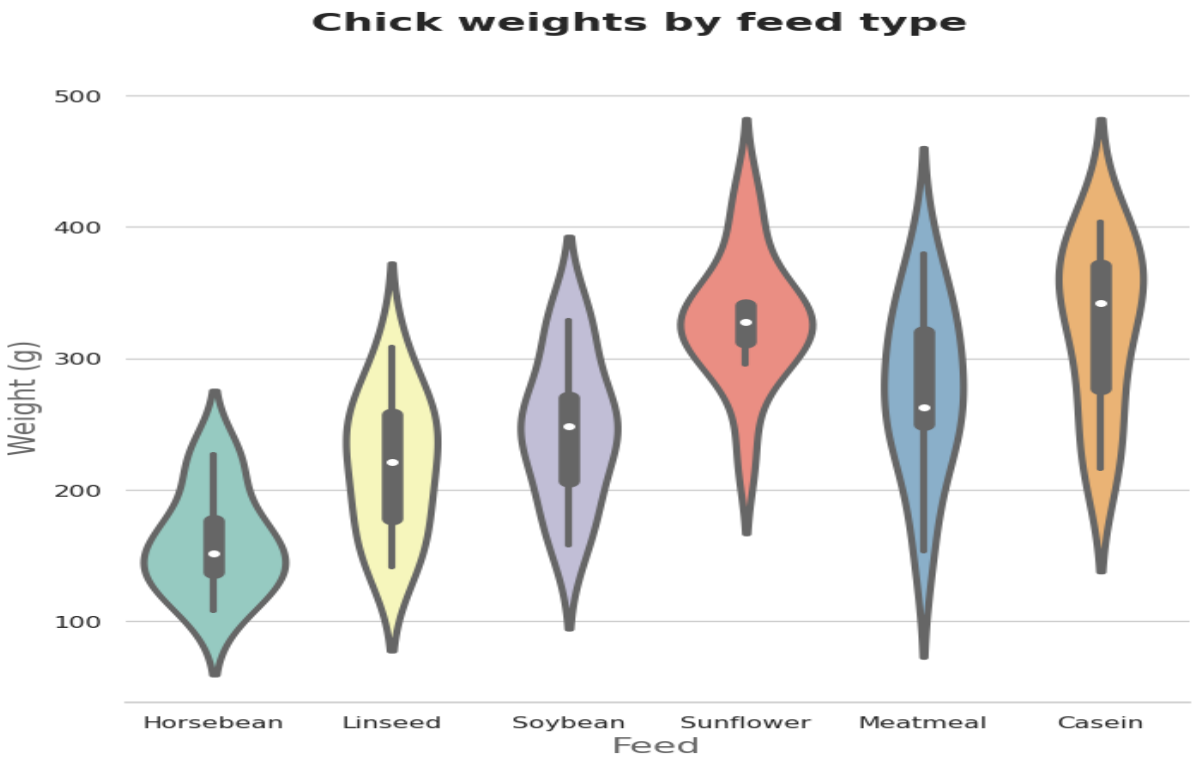
A violin plot is a hybrid of a box plot and a kernel density plot, which shows peaks in the data, distributions of numeric data for one or more groups using density curves. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable. The width of each curve corresponds with the approximate frequency of data points in each region.

Violin plots are used when you want to observe the distribution of numeric data and are especially useful when you want to make a comparison of distributions between multiple groups. The peaks, valleys, and tails of each group's density curve can be compared to see where groups are similar or different. Additional elements, like box plot quartiles, are often added to a violin plot to provide additional ways of comparing groups

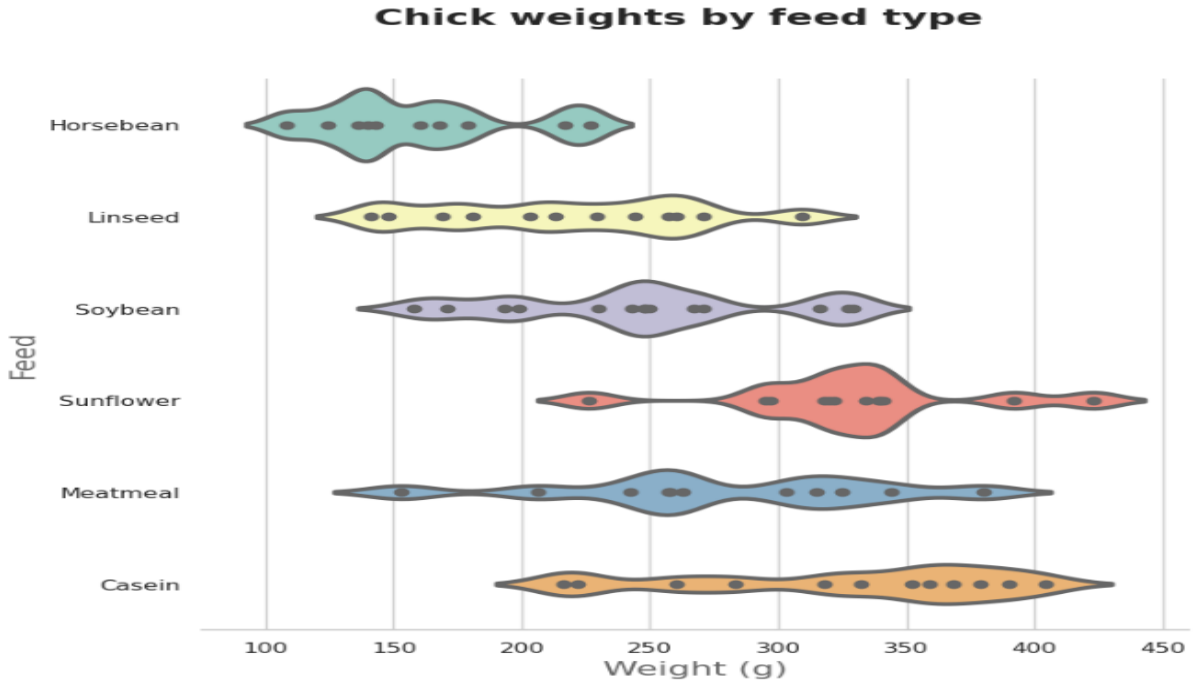


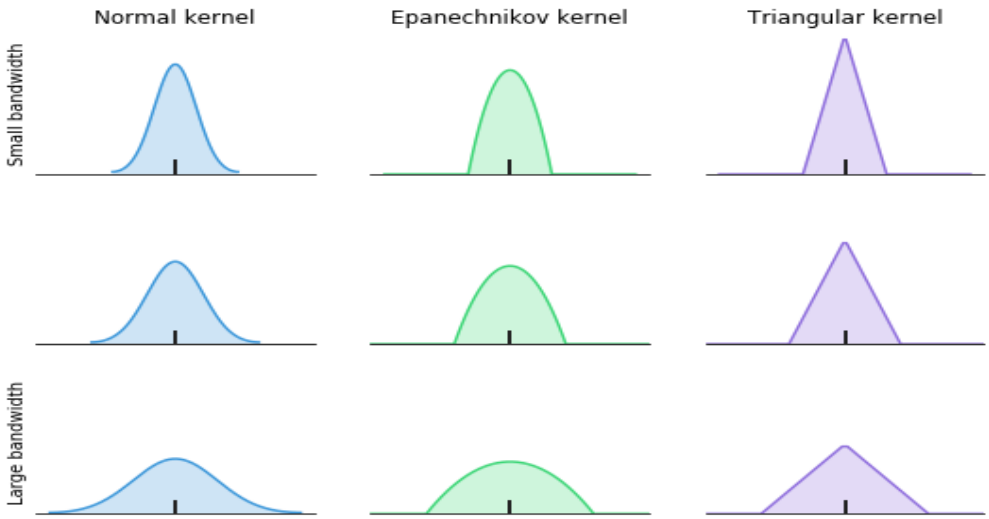
The violin plot was formally introduced by Hintze and Nelson (1998) as an adaptation of Tukey's (1977) box plot in combination with density traces. In their development of the violin plot, Hintze and Nelson built on previous work by Benjamini (1988) who suggested opening the box of a boxplot to "convey information about the density of the

values in a batch" (p. 257). One variation of the boxplot introduced by Benjamini was the vaseplot: "A boxplot where the width of the box at each point is proportional to the estimated density there" (p. 259). Thus, the vaseplot and its successor the violin plot does not depict raw data, but rather the estimated density of scores that fall within a given interval. The width of the violin plot on each side may be interpreted as a smoothed histogram of data density (Hu, 2020). The advantage of this is that the graph remains readable with the possibility to include graphical descriptions of summaries as well (Benjamini, 1988). In addition, it is possible to add information about the raw data by jittering data points randomly along the x-axis to avoid cluttering (Benjamini, 1988).



This violin plot shows the relationship of feed type to chick weight. The box plot elements show the median weight for horsebean-fed chicks is lower than for other feed types. The shape of the distribution (extremely skinny on each end and wide in the middle) indicates the weights of sunflower-fed chicks are highly concentrated around the median.





Sometimes the median and mean aren't enough to understand a dataset. Are most of the values clustered around the median? Or are they clustered around the minimum and the maximum with nothing in the middle? When you have questions like these, distribution plots are your friends.

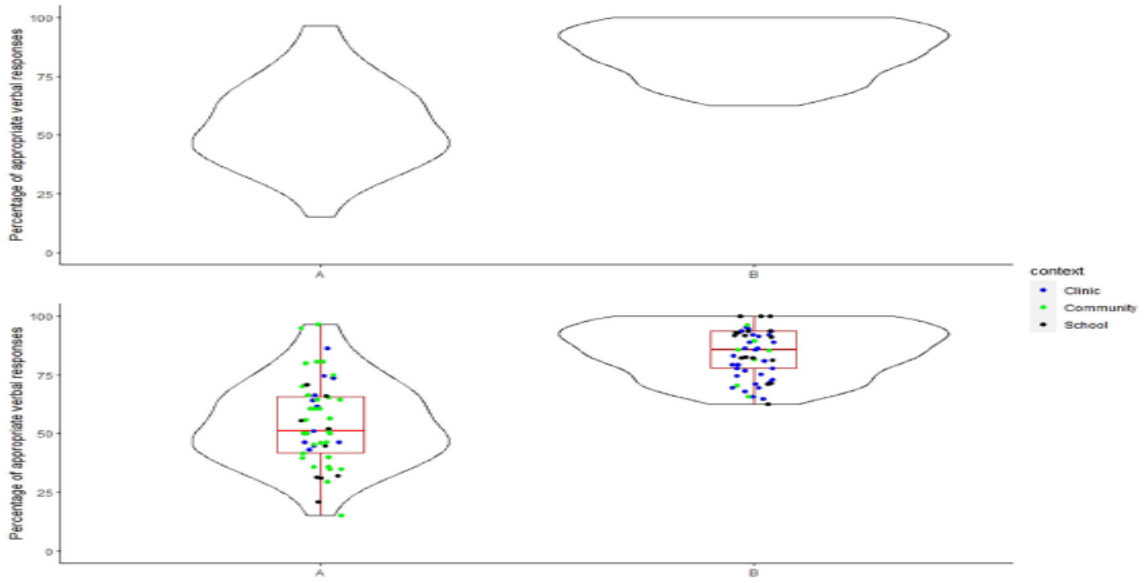
The box plot is an old standby for visualizing basic distributions. It's convenient for comparing summary statistics (such as range and quartiles), but it doesn't let you see variations in the data. For multimodal distributions (those with multiple peaks) this can be particularly limiting this is where the violin plot comes into picture.

The density curve, aka kernel density plot or kernel density estimate (KDE), is a less-frequently encountered depiction of data distribution, compared to the more common histogram. Below, we'll perform a brief explanation of how density curves are built.

In a KDE, each data point contributes a small area around its true value. The shape of this area is called the kernel function. Kernels can take different shapes from smooth bell curves to sharp triangular peaks. In addition, kernels can have different width, or bandwidth, affecting the influence of each individual data point. Bandwidth size is usually determined by using mathematical rules of thumb but can be tweaked depending on the shape and skew of the data to be plotted.

Figure 1

Violin Plots for the Koegel et al. (1992) Data for Participant Tony



Note. The upper panel shows the basic violin plots. The lower panel shows the same violin plots with the addition of a boxplot and jitter of the raw scores. Blue indicates clinical setting, green indicates community setting, and black indicates school setting.

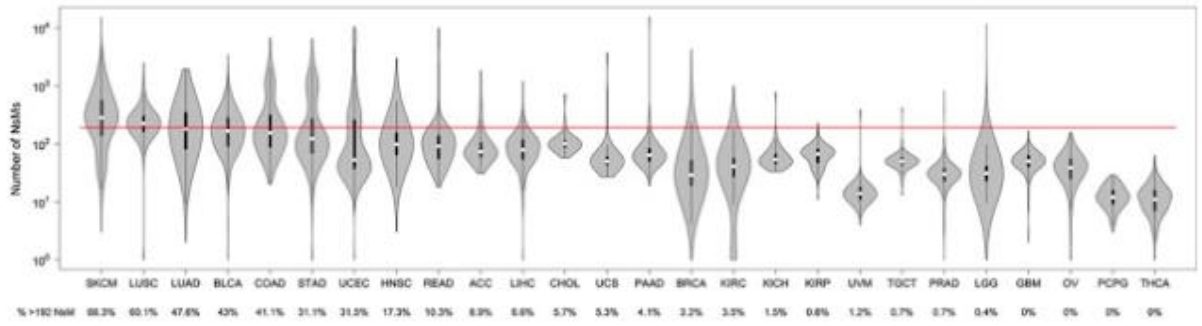
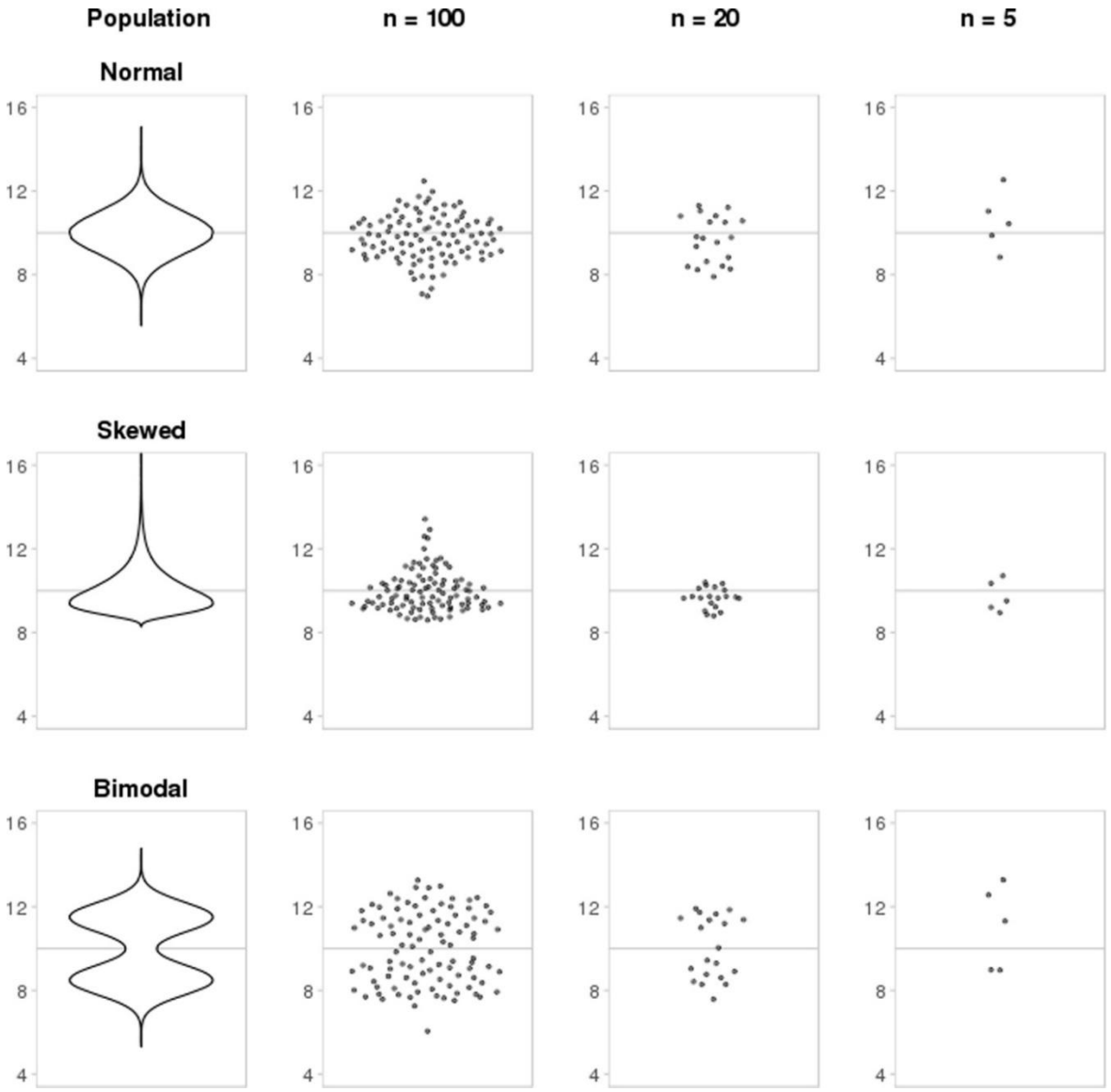
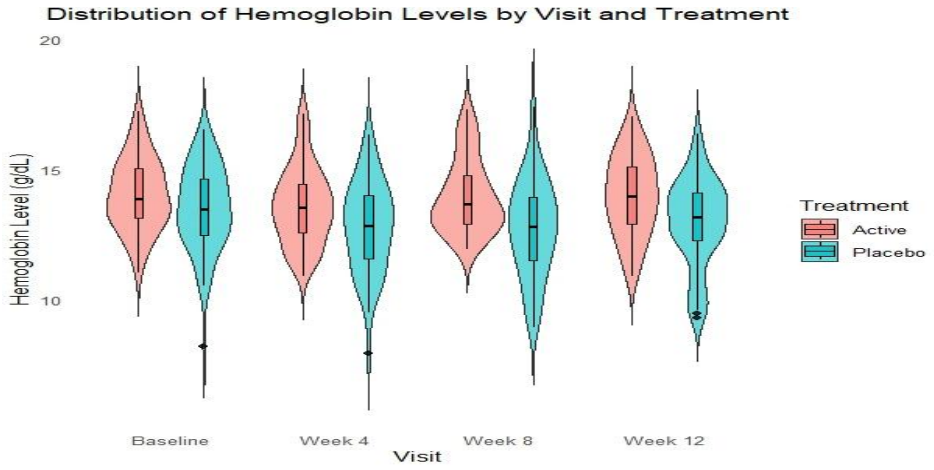


Figure 2.
Violin plot for the number of NsM (log10) across 26 tumor types in TCGA data. Red line is the cutoff for 192 NsM. ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; LLG, brain lower grade glioma; BRCA, breast invasive carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; UCS, uterine carcinosarcoma; UCEC, uterine corpus endometrial carcinoma; and UVM, uveal melanoma.

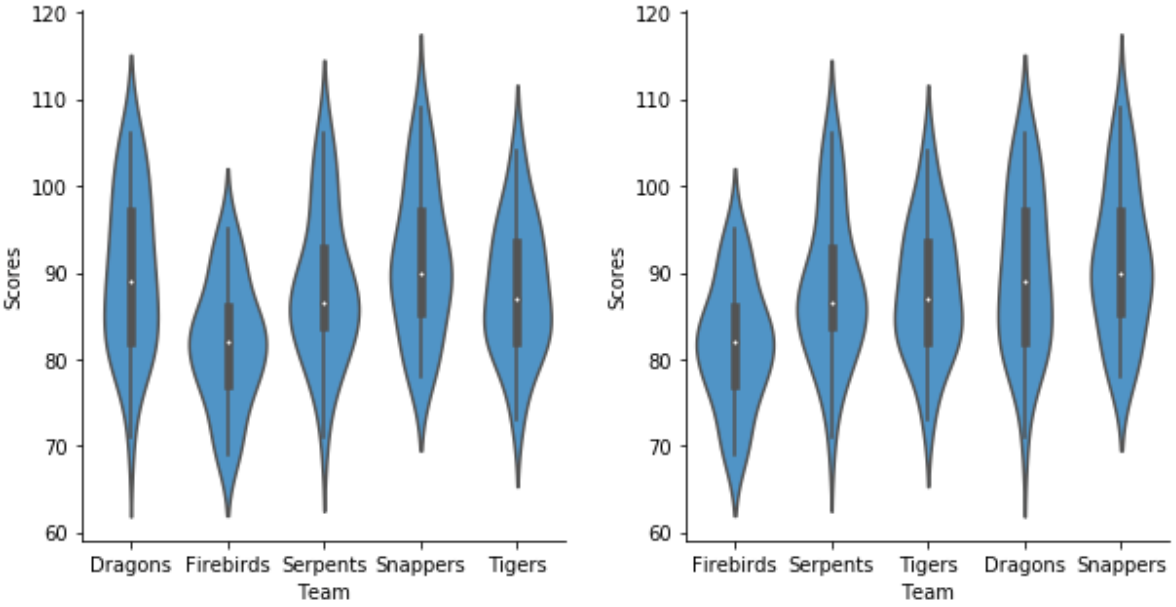


Let us look at some other examples of Violin Plots to understand how it can be used for displaying different types of data.

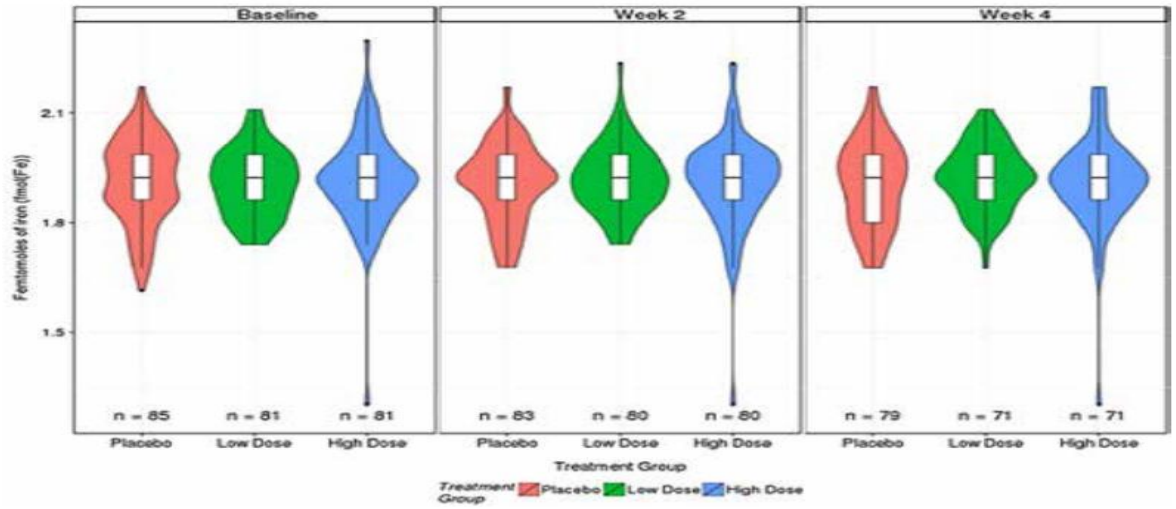
Example 1



Example 2



Example 3:



Data Structure for Violin Plot

Let us look at the example data set for creating a violin plot. Each row corresponds with a single data point, while cell values indicate group membership and numeric value for each point. All the plot features will be automatically calculated from this raw input. If all the data is in a single group, then the column indicating group membership will not be necessary.

Data Interpretation

- **CR (Complete Response)**: 100% shrinkage
- **PR (Partial Response)**: More than 30% shrinkage
- **SD (Stable Disease)**: Between -30% and +20%
- **PD (Progressive Disease)**: Growth beyond 20%

Subject ID	Treatment	Biomarker Level	Tumor Response (mm)	RECIST Category
10001	Drug A	23.5	-15	PR (Partial Response)
10002	Placebo	18.2	+10	PD (Progressive Disease)
10003	Drug A	27.8	-25	PR (Partial Response)
10004	Drug A	19.4	-5	SD (Stable Disease)
10005	Placebo	16.5	+12	PD (Progressive Disease)
10006	Drug A	30.1	-30	CR (Complete Response)
10007	Placebo	21.2	0	SD (Stable Disease)
10008	Drug A	25.3	-20	PR (Partial Response)
10009	Placebo	17.8	+8	PD (Progressive Disease)
10010	Drug A	29.5	-35	CR (Complete Response)

Violon Plot using R Studio

```r
# Load library
library(ggplot2)
library(dplyr)
library(gridExtra)

# Set seed for reproducibility
set.seed(123)

# Create Tumor response & Biomarker dataset
oncology_data <- data.frame(
  Subject_ID = sprintf("%05d", 1:100),
  Treatment = rep(c("Drug A", "Placebo"), each = 50),
  Tumor_Response = c(rnorm(50, mean = -65, sd = 20),
                     rnorm(50, mean = -20, sd = 25)),
  Biomarker_Level = c(rnorm(50, mean = 15, sd = 5),
                      rnorm(50, mean = 20, sd = 6))
)

# Assign RECIST categories based on tumor response percentage
oncology_data <- oncology_data %>%
  mutate(RECIST = case_when(
    Tumor_Response <= -80 ~ "CR",   # Complete Response
    Tumor_Response <= -30 ~ "PR",    # Partial Response
    Tumor_Response > -30 & Tumor_Response < 20 ~ "SD",  # Stable Disease
    Tumor_Response >= 20 ~ "PD"     # Progressive Disease
  ))

# Display first 15 rows
head(oncology_data,15)
```

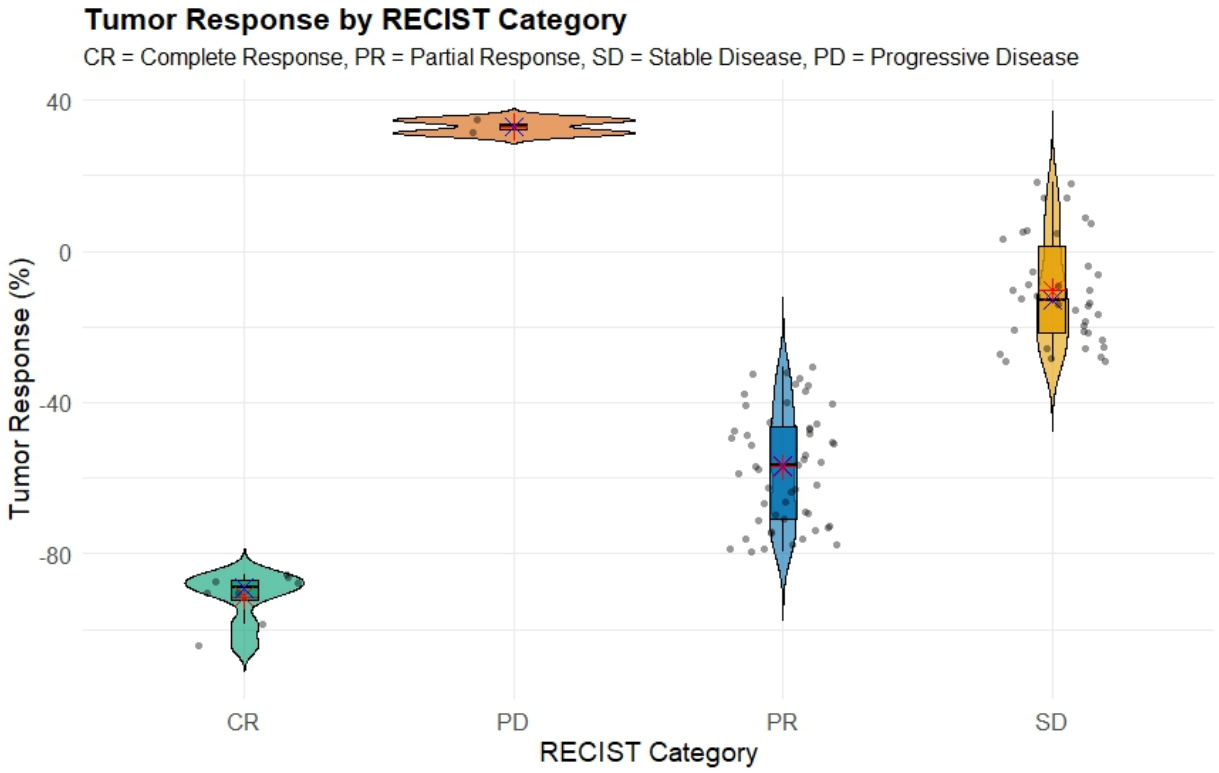
```
> # Display first 15 rows
> head(oncology_data,15)
   Subject_ID Treatment Tumor_Response Biomarker_Level RECIST
1       00001    Drug A       -76.20951        11.447967     PR
2       00002    Drug A       -69.60355        16.284419     PR
3       00003    Drug A       -33.82583        13.766541     PR
4       00004    Drug A       -63.58983        13.262287     PR
5       00005    Drug A       -62.41425        10.241907     PR
6       00006    Drug A       -30.69870        14.774861     PR
7       00007    Drug A       -55.78168        11.075478     PR
8       00008    Drug A       -90.30122         6.660290     CR
9       00009    Drug A       -78.73706        13.098867     PR
10      00010    Drug A       -73.91324        19.594983     PR
11      00011    Drug A       -40.51836        12.123265     PR
12      00012    Drug A       -57.80372        18.039822     PR
13      00013    Drug A       -56.98457         6.910586     PR
14      00014    Drug A       -62.78635        14.722190     PR
15      00015    Drug A       -76.11682        17.597036     PR
```

Violin plots allow researchers to compare the distribution of tumor reduction and biomarker levels across different treatment groups.

Example 1: Violin Plot: Tumor Response (%) by RECIST Category

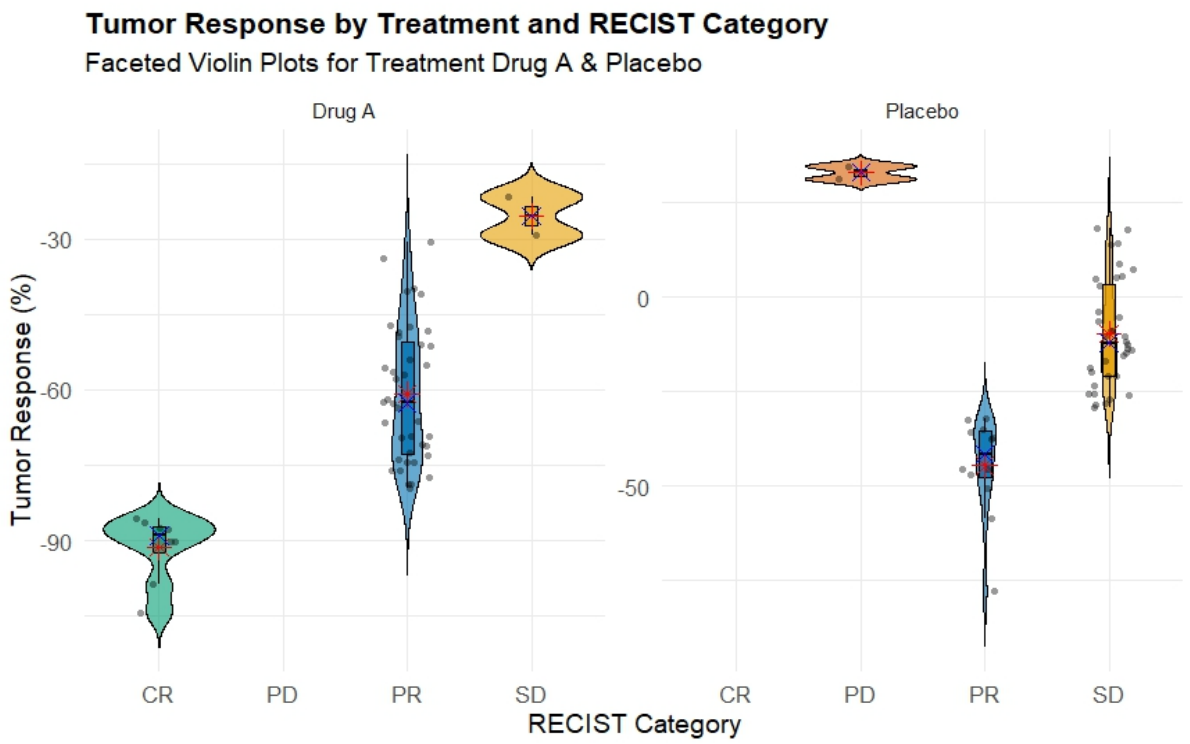
```
# Custom color scheme for RECIST
recist_colors <- c("CR" = "#009E73", "PR" = "#0072B2", "SD" = "#E69F00", "PD" = "#D55E00")

# Violin plot by RECIST category
ggplot(oncology_data, aes(x = RECIST, y = Tumor_Response, fill = RECIST)) +
  geom_violin(trim = FALSE, alpha = 0.6, color = "black") +
  geom_boxplot(width = 0.1, color = "black", alpha = 0.8, outlier.shape = NA) +
  geom_jitter(shape = 16, position = position_jitter(0.2), alpha = 0.4, color = "black") +
  stat_summary(fun = "mean", geom = "point", shape = 8, size = 4, color = "red") +
  stat_summary(fun = "median", geom = "point", shape = 4, size = 4, color = "blue") +
  scale_fill_manual(values = recist_colors) +
  labs(title = "Tumor Response by RECIST Category",
       subtitle = "CR = Complete Response, PR = Partial Response, SD = Stable Disease, PD = Progressive Disease",
       x = "RECIST Category",
       y = "Tumor Response (%)") +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", size = 16),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```



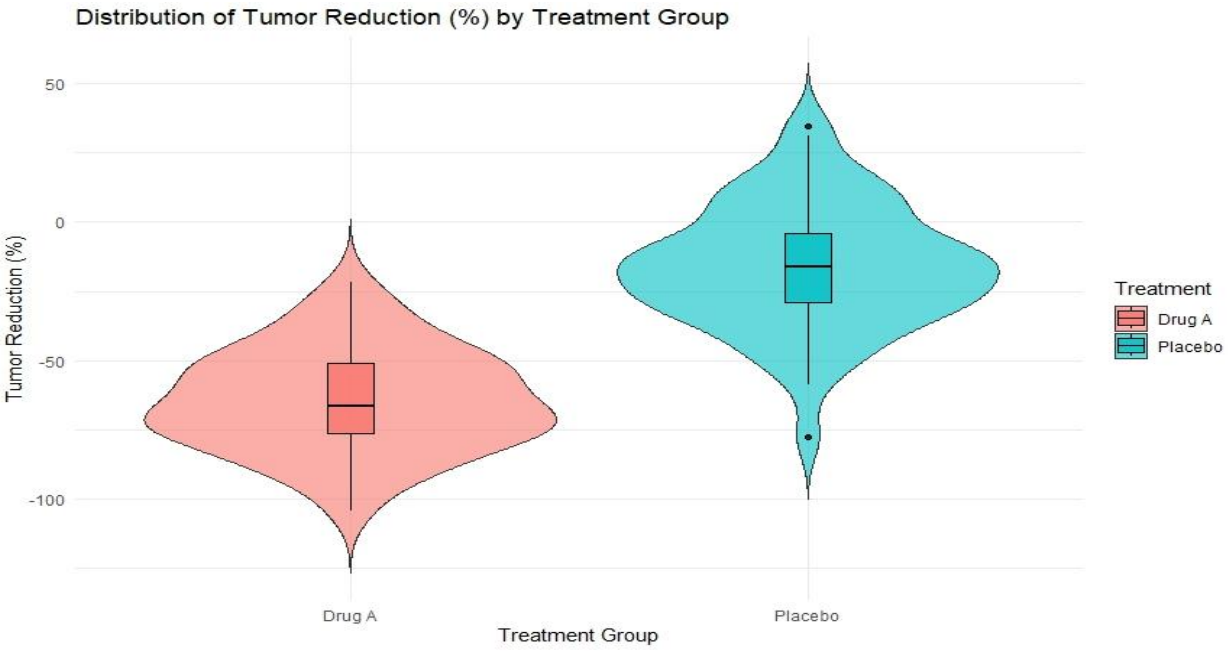
Example 2: Faceted Violin Plot: Tumor Response by Treatment & RECIST

```
# Violin Plot with facet wrap for Treatment and RECIST category
ggplot(oncology_data, aes(x = RECIST, y = Tumor_Response, fill = RECIST)) +
  geom_violin(trim = FALSE, alpha = 0.6, color = "black") +
  geom_boxplot(width = 0.1, color = "black", alpha = 0.8, outlier.shape = NA) +
  geom_jitter(shape = 16, position = position_jitter(0.2), alpha = 0.4, color = "black") +
  stat_summary(fun = "mean", geom = "point", shape = 8, size = 4, color = "red") +
  stat_summary(fun = "median", geom = "point", shape = 4, size = 4, color = "blue") +
  scale_fill_manual(values = recist_colors) +
  labs(title = "Tumor Response by Treatment and RECIST Category",
       subtitle = "Faceted Violin Plots for Treatment Drug A & Placebo",
       x = "RECIST Category",
       y = "Tumor Response (%)") +
  facet_wrap(~Treatment, scales = "free_y") +  # Facet by Treatment
  theme_minimal(base_size = 14) +
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", size = 16),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```



Example 3: Violin Plot for Tumor Reduction by Treatment Group

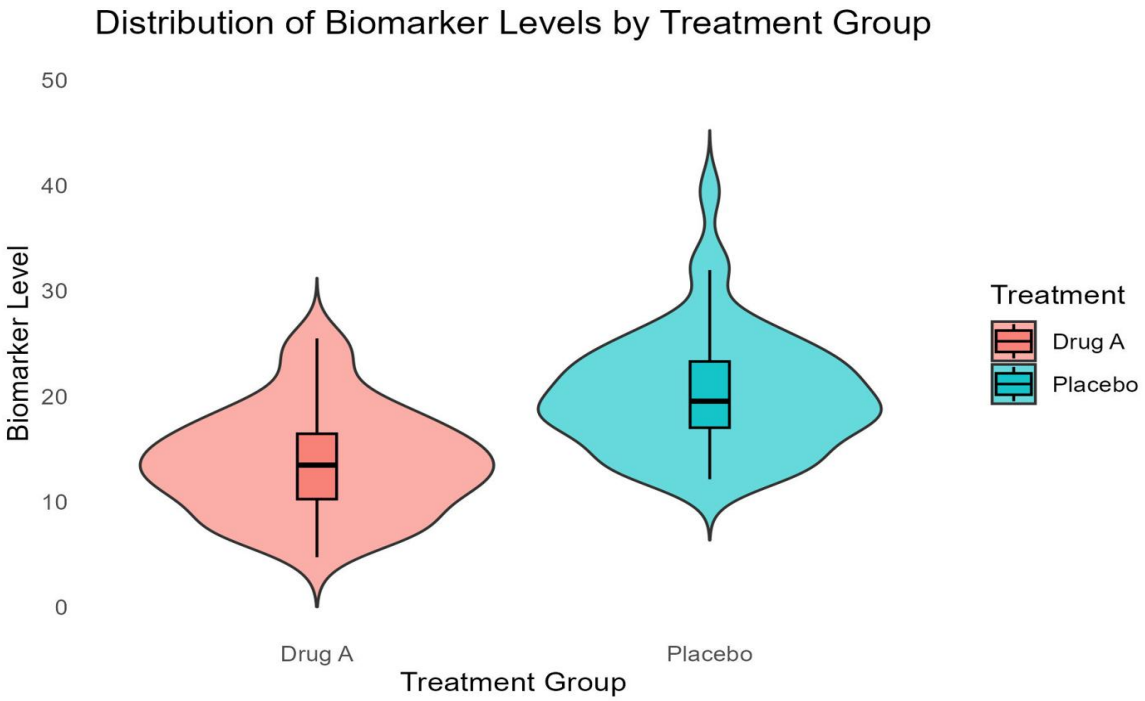
```r
# Violin Plot for Tumor Reduction
ggplot(oncology_data, aes(x = Treatment, y = Tumor_Response, fill = Treatment)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, color = "black", alpha = 0.8) +
  labs(title = "Distribution of Tumor Reduction (%) by Treatment Group",
       x = "Treatment Group",
       y = "Tumor Reduction (%)") +
  theme_minimal()
```



Tumor Reduction: The violin plot illustrates how tumor reduction varies between treatments Drug A and Placebo. Drug A appears to have a higher median tumor reduction with a wider distribution.

Example 4: Violin Plot for Biomarker Levels by Treatment Group

```
# Violin Plot for Biomarker Levels
ggplot(oncology_data, aes(x = Treatment, y = Biomarker_Level, fill = Treatment)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, color = "black", alpha = 0.8) +
  labs(title = "Distribution of Biomarker Levels by Treatment Group",
       x = "Treatment Group",
       y = "Biomarker Level") +
       ylim(0,50)+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))  # Center-align title
```



Biomarker Levels: The biomarker levels differ significantly between treatment groups, indicating a possible correlation with treatment efficacy.

Exploring various Output Formats

A few other output formats like RTF, PNG, JPEG etc. can be used to save graph as per requirements.

```
# Save output to PDF or jpeg or png or tiff format
ggsave(filename = "C:/Users/MrityunjayKumar/Desktop/US Connect 2025/Violin_plot.pdf",
       plot = violin, device = "pdf", width = 10, height = 6)

ggsave(filename = "C:/Users/MrityunjayKumar/Desktop/US Connect 2025/Violin_plot.jpeg",
       plot = violin, device = "jpeg", width = 6, height = 4)

ggsave(filename = "C:/Users/MrityunjayKumar/Desktop/US Connect 2025/Violin_plot.png",
       plot = violin, device = "png", width = 6, height = 4)
```

Comparison with alternative visualisation methods (Box Plot, Histogram and Density Plot)

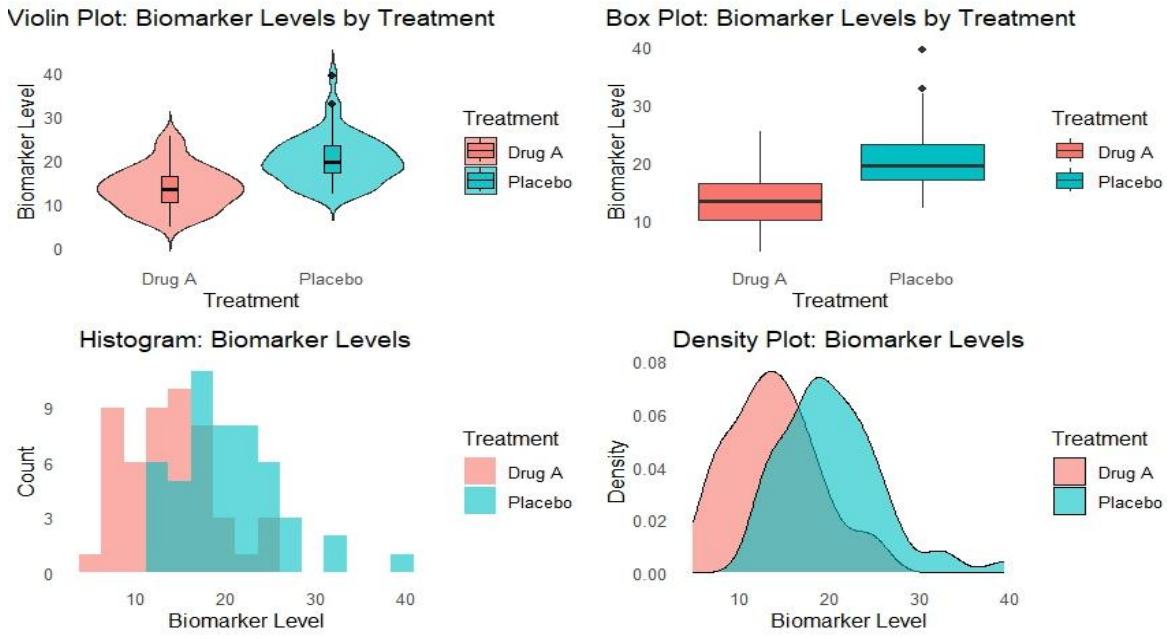
```
# Violin Plot
violin_plot <- ggplot(oncology_data, aes(x = Treatment, y = Biomarker_Level, fill = Treatment)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, color = "black", alpha = 0.8) +
  labs(title = "Violin Plot: Biomarker Levels by Treatment", x = "Treatment", y = "Biomarker Level") +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  ) +
  theme(plot.title = element_text(hjust = 0.5))

# Box Plot
box_plot <- ggplot(oncology_data, aes(x = Treatment, y = Biomarker_Level, fill = Treatment)) +
  geom_boxplot() +
  labs(title = "Box Plot: Biomarker Levels by Treatment", x = "Treatment", y = "Biomarker Level") +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```

```
# Histogram
histogram_plot <- ggplot(oncology_data, aes(x = Biomarker_Level, fill = Treatment)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 15) +
  labs(title = "Histogram: Biomarker Levels", x = "Biomarker Level", y = "Count") +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  ) +
  theme(plot.title = element_text(hjust = 0.5))

# Density Plot
density_plot <- ggplot(oncology_data, aes(x = Biomarker_Level, fill = Treatment)) +
  geom_density(alpha = 0.6) +
  labs(title = "Density Plot: Biomarker Levels", x = "Biomarker Level", y = "Density") +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  ) +
  theme(plot.title = element_text(hjust = 0.5))

# Arrange all plots in a grid
grid.arrange(violin_plot, box_plot, histogram_plot, density_plot, ncol = 2)
```



Discussion

How violin plots enhance the understanding of oncology data

Violin plots go beyond traditional boxplots by providing a richer view of oncology data distributions, helping researchers detect patterns, identify variability, and make more informed decisions in clinical and pharmaceutical studies.

Violin plots provide a comprehensive visualisation of oncology data by combining distribution shape and summary statistics in a single plot. They help identify subpopulations, compare treatment responses, and detect data skewness and outliers effectively.

Unlike box plots or histograms, violin plots retain full data distribution while highlighting density variations, making them useful for assessing biomarker levels and tumor response trends.

Their ability to present asymmetry, multiple peaks, and variability enhances clinical decision-making, offering deeper insights into treatment effects in oncology trials.

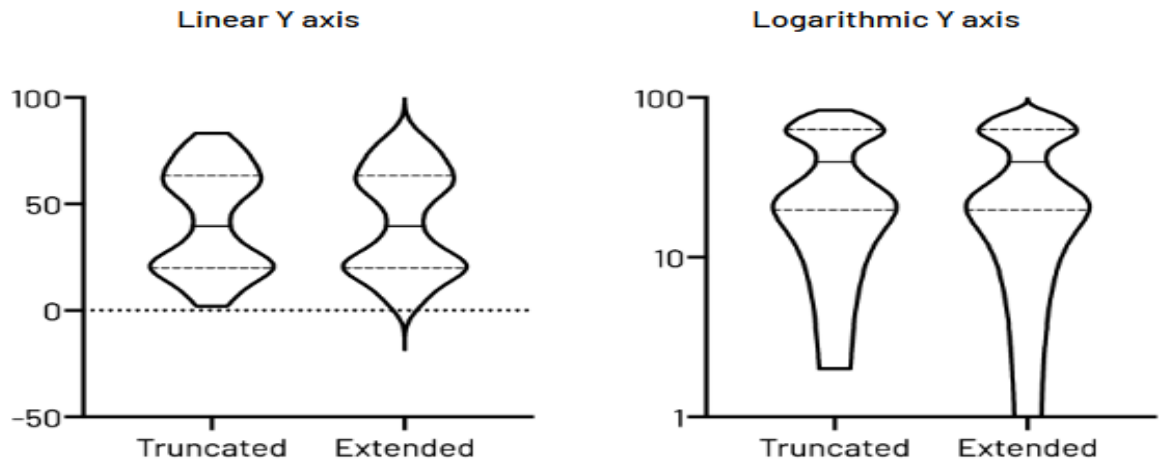
Usage	How Violin Plots Help in Oncology Data Analysis
Data Distribution	Displays the full probability distribution, showing if biomarker levels or tumor responses follow a normal or skewed distribution.
Variability Assessment	Highlights the spread of data within each treatment group, indicating whether responses are consistent or highly variable.
Multimodal Distributions	Helps detect subpopulations in oncology data by revealing multiple peaks that standard boxplots might miss.
Comparing Treatment Responses	Enables visual comparison of different treatments by showing how their distributions differ in shape and spread.
Outlier Identification	Provides better context for outliers by showing density, which is useful in identifying exceptional responders or adverse effects.
Clinical Interpretation	Helps researchers and clinicians detect shifts in distributions, indicating treatment efficacy or toxicity concerns.
Oncology	Comparing biomarker levels (e.g., PD-L1, VEGF) across treatment arms
	Evaluating tumor shrinkage over time
	Representing survival analysis (PFS, OS) distributions

Limitations of Violin Plot

When considering a violin plot that has been graphed on a logarithmic Y axis, there are two important issues that must be considered. Each of these two issues result in their own unique visual properties of the violin plots (when using a logarithmic axis), and each can lead to serious confusion if not handled properly. A summary of these two issues is illustrated as follows:

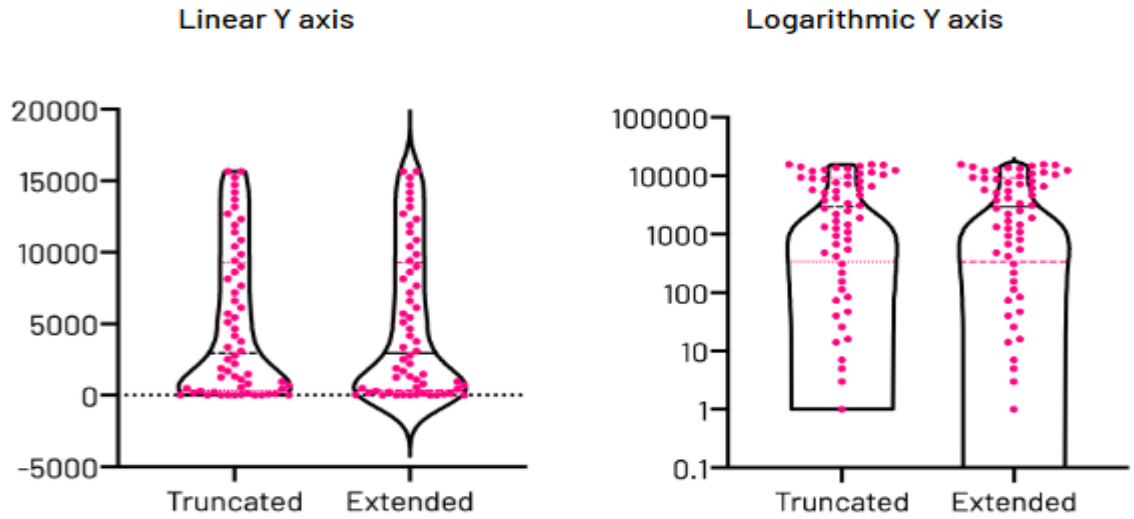
Issue # 1

Even though the data used to generate a violin plot contains only positive numbers, the violin itself may extend beyond zero into negative values. This is problematic because logarithms can't be negative (or zero).



Issue # 2

The width of violin plots is determined by examining the distance between values in a linear fashion. This is problematic because the distance between values on a logarithmic axis is not uniform.



Conclusion

Violin plots provide a powerful way to visualise complex clinical data distributions in oncology research. By incorporating these visualizations into statistical programming workflows, statisticians/programmers can gain deeper insights into treatment effects and biomarker variations.

The future of data science lies in reproducible, robust methods that communicate our results to a larger community. It is expected that Violin plot may help you to better understand and communicate your own data insights into a more meaningful way. We have highlighted some of the strengths of these plots compared to traditional methods such as bar and box plots which adds a step further in visualisation of multimodal data.

References

Violin Plots as Visual Tools in Meta-Analysis: https://meth.psychopen.eu/index.php/meth/article/view/9209/9209.html

https://pmc.ncbi.nlm.nih.gov/articles/PMC8947810/

https://www.atlassian.com/data/charts/violin-plot-complete-guide#:~:text=Violin%20plots%20are%20used%20when,groups%20are%20similar%20or%20different.

https://onbiostatistics.blogspot.com/2023/05/violin-plot-versus-box-whisker-plot.html

https://www.fda.gov/media/96653/download

https://www.graphpad.com/support/faq/violin-plots-and-logarithmic-axes/

Acknowledgments

Authors would like to extend their sincere thanks to Ephicacy Lifescience Analytics for giving them an opportunity to write and present this paper. Any brand and product names are trademarks of their respective companies.

Recommended Reading

- *https://mode.com/blog/violin-plot-examples*
- *https://github.com/r2evans/violinplot*

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Mrityunjay Kumar, Associate Director II - Statistical Programming
Ephicacy Lifescience Analytics
Bengaluru, India
E-mail: mrityunjay.kumar@ephicacy.com
www.ephicacy.com
https://www.linkedin.com/in/mrityunjay-kumar-b8aa3917/

Shashikant Kumar, Director - Statistical Programming
Ephicacy Lifescience Analytics
Bengaluru, India
E-mail: shashikant.kumar@ephicacy.com
www.ephicacy.com