EXPT NO:6 DATE: 09.02.2026	OVER-PLOTTING REDUCTION TECHNIQUES

PRE-LAB QUESTIONS

1. **Why is over-plotting common in big data visualization?**
   Over-plotting is common because big data contains thousands or millions of points, and many points fall in the same positions on a graph. This causes points to overlap and hide the actual distribution.

2. **How does data density affect perception?**
   When data density is high, the plot looks like a dark clustered region. This makes it difficult to identify important patterns like clusters, trends, and outliers, reducing the readability of the visualization.

3. **What trade-offs exist between detail and clarity?**
   - Showing every data point gives high detail, but it creates clutter and reduces clarity.
   - Using aggregation/binning improves clarity, but it reduces the visibility of individual data points.
     So the trade-off is between full detail vs clear understanding.

4. **How do AI datasets increase visualization complexity?**
   AI datasets are usually large-scale, high-dimensional, and continuous, with many features and classes. This increases overlap in plots and makes it harder to visualize relationships clearly using normal plotting methods.

5. **Why is over-plotting a serious analytical risk?**
   Over-plotting can hide important information such as:
   - minority patterns
   - outliers
   - true clusters
   - real correlations

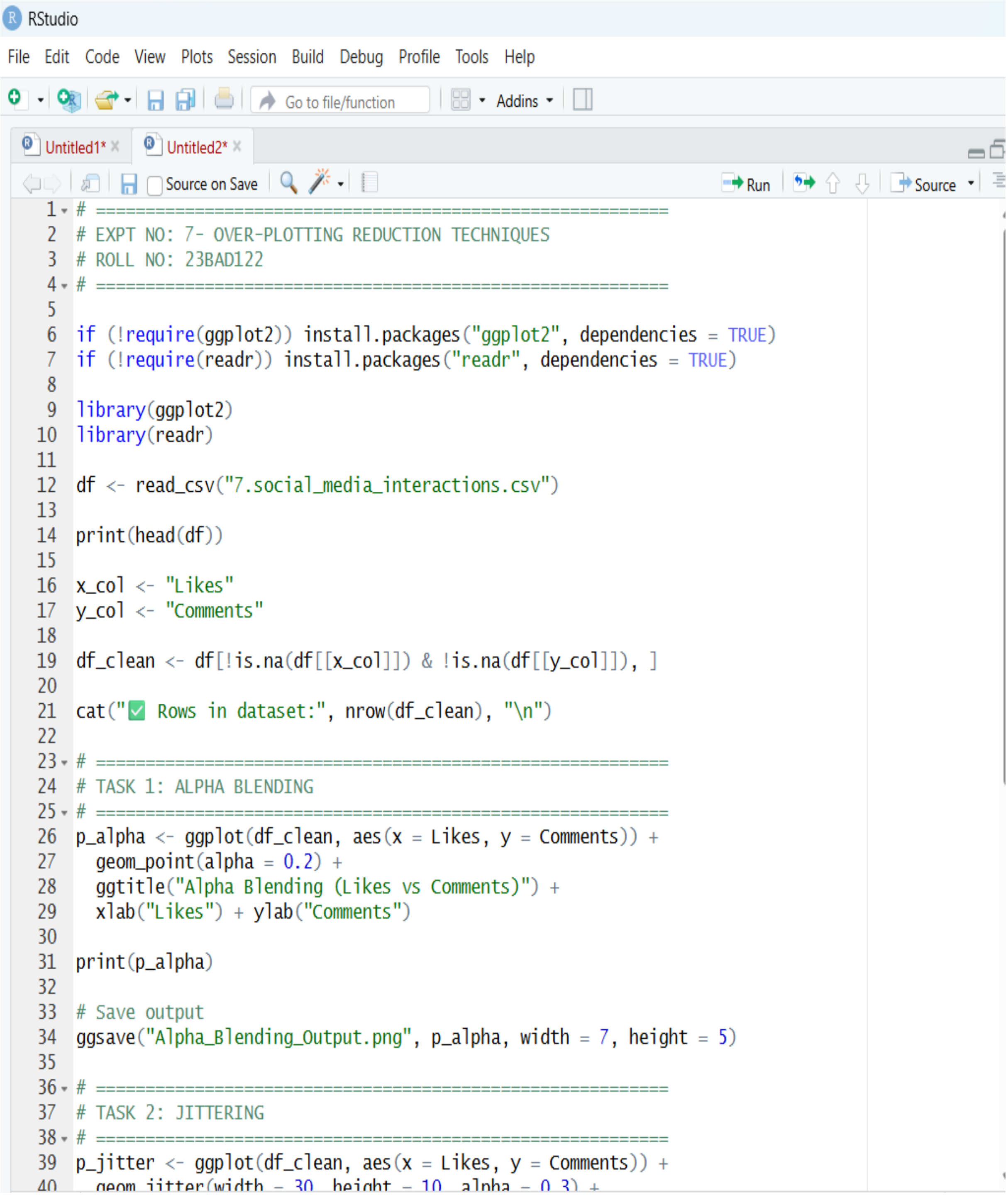
   This can lead to wrong conclusions and wrong decisions, especially in AI and business analytics.

OBJECTIVE : To apply techniques that reduce visual clutter in large-scale datasets.

SCENARIO A social media analytics company visualizes millions of user interactions to study engagement patterns.

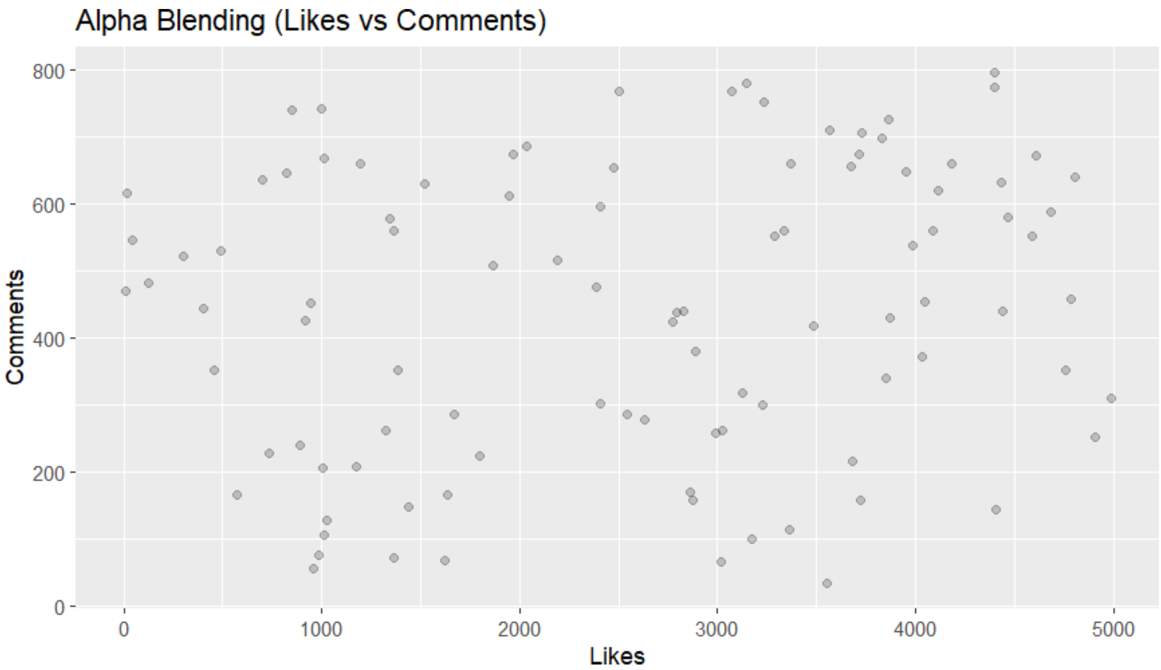
IN-LAB TASKS (Using R Language) • Apply alpha blending • Implement jittering techniques • Use aggregation and binning

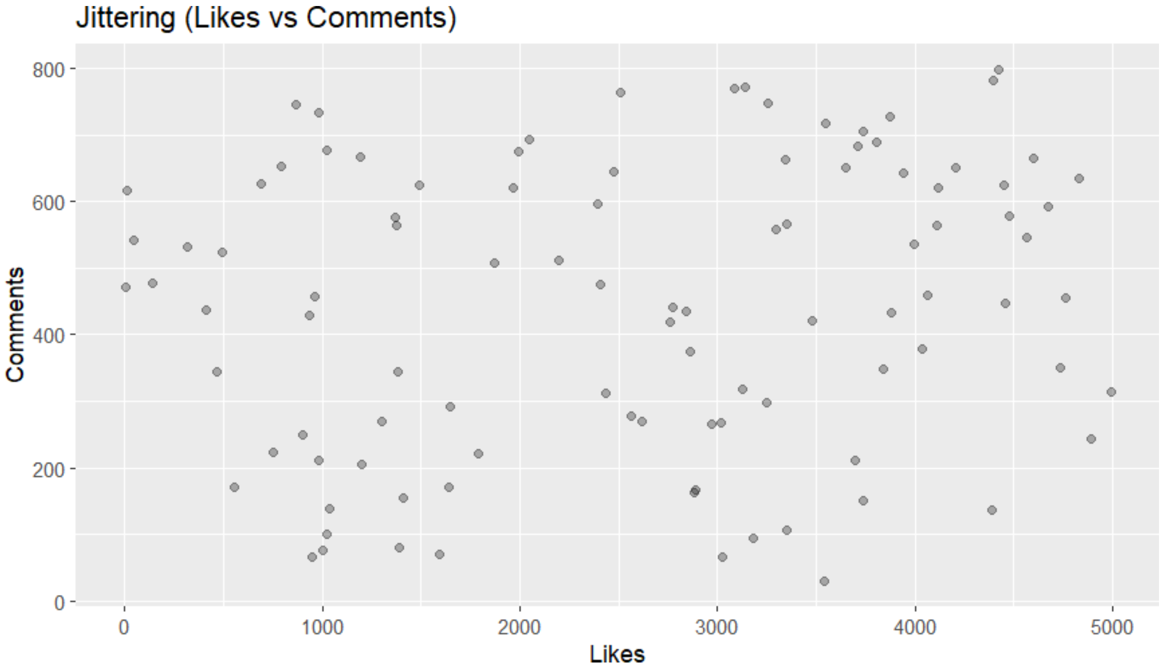
CODE:

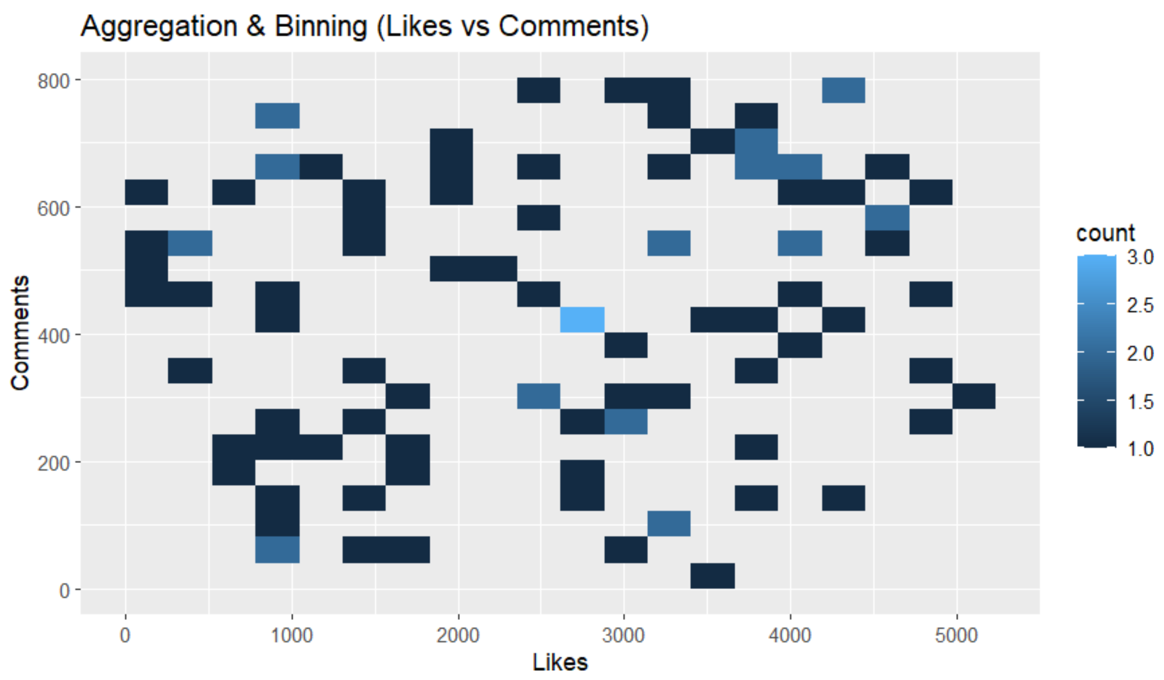


```
# ==========================================================
# EXPT NO: 7- OVER-PLOTTING REDUCTION TECHNIQUES
# ROLL NO: 23BAD122
# ==========================================================

if (!require(ggplot2)) install.packages("ggplot2", dependencies = TRUE)
if (!require(readr)) install.packages("readr", dependencies = TRUE)

library(ggplot2)
library(readr)

df <- read_csv("7.social_media_interactions.csv")

print(head(df))

x_col <- "Likes"
y_col <- "Comments"

df_clean <- df[!is.na(df[[x_col]]) & !is.na(df[[y_col]]), ]

cat("✅ Rows in dataset:", nrow(df_clean), "\n")

# ==========================================================
# TASK 1: ALPHA BLENDING
# ==========================================================
p_alpha <- ggplot(df_clean, aes(x = Likes, y = Comments)) +
  geom_point(alpha = 0.2) +
  ggtitle("Alpha Blending (Likes vs Comments)") +
  xlab("Likes") + ylab("Comments")

print(p_alpha)

# Save output
ggsave("Alpha_Blending_Output.png", p_alpha, width = 7, height = 5)

# ==========================================================
# TASK 2: JITTERING
# ==========================================================
p_jitter <- ggplot(df_clean, aes(x = Likes, y = Comments)) +
  geom_jitter(width = 30, height = 10, alpha = 0.3) +
```

OUTPUT:


Alpha Blending (Likes vs Comments)
Comments
Likes
0
200
400
600
800
0
1000
2000
3000
4000
5000


Jittering (Likes vs Comments)
Comments
Likes
0
200
400
600
800
0
1000
2000
3000
4000
5000



POST-LAB QUESTIONS

1. **Which technique provided the best clarity and why?**
   Aggregation and binning (2D bin / hexbin) provided the best clarity because it groups overlapping points into bins and shows data density clearly. This makes patterns visible even when the dataset is large.

2. **How does over-plotting distort analytical conclusions?**
   Over-plotting hides important patterns such as:
   - clusters
   - trends
   - outliers
   - rare behaviors

   This can make the data look misleading and can cause incorrect conclusions about relationships between variables.

3. **When should aggregation be preferred over raw plotting?**
   Aggregation should be preferred when:
   - the dataset is very large
   - points overlap heavily
   - the goal is to understand overall patterns (macro trends)
   - raw scatter plots become unreadable

4. **How do these techniques support scalable AI analytics?**
   These techniques help scalable AI analytics by:
   - revealing real patterns in large datasets
   - improving readability and interpretability
   - supporting better feature analysis
   - helping detect bias, imbalance, and outliers
   - making large AI datasets easier to validate visually

5. **Explain real-world consequences of ignoring over-plotting.**
   Ignoring over-plotting can cause:
   - wrong business decisions (marketing, targeting, engagement)
   - missing fraud/spam patterns
   - incorrect trend detection
   - poor AI model training insights
   - failure to identify rare but important user behaviors

LEARNING OUTCOME: Students master over-plotting reduction for big data visual analytics.

ASSESSMENT

Description	**Max Marks**	**Marks Awarded**
Pre Lab Exercise	**5**	
In Lab Exercise	**10**	
Post Lab Exercise	**5**	
Viva	**10**	
Total	**30**	
	Faculty Signature	