3_2023900021

hrithik.rachakonda@research.iiit.ac.in

2026-02-02

3 Data Plotting Adventure

Question 1:

3.1 Subtask 1: The Last of Us You're part of a research team analyzing the survival outcomes of individuals in different types of locations during a zombie apocalypse. The goal is to understand where survival chances are highest and what the common outcomes are in each location. • 118 people from the Safe Zone, who were males, turned into zombies. • 62 people from the Safe Zone, who were males, survived. • 4 people from the Safe Zone, who were females, turned into zombies. • 141 people from the Safe Zone, who were females, survived. • 154 people from the Contaminated City, who were males, turned into zombies. • 25 people from the Contaminated City, who were males, survived. • 13 people from the Contaminated City, who were females, turned into zombies. • 93 people from the Contaminated City, who were females, survived. • 422 people from the Rural Area, who were males, turned into zombies. • 88 people from the Rural Area, who were males, survived. 2 • 106 people from the Rural Area, who were females, turned into zombies. • 90 people from the Rural Area, who were females, survived. • 670 people from the Isolated Island, who were males, turned into zombies. • 192 people from the Isolated Island, who were males, survived. • 3 people from the Isolated Island, who were females, turned into zombies. • 20 people from the Isolated Island, who were females, survived.

Answer 1:

Visualizations used

1. **Stacked proportion bar chart (100% stacked), faceted by Gender**
2. **Stacked count bar chart, faceted by Gender**

Why these are the right choices

- The **proportion plot** directly answers "**survival chances are highest where?**" because it compares survival *rates* (not affected by different total group sizes).
- The **count plot** answers "**what common outcomes occur most?**" because it shows absolute numbers (useful for understanding burden/scale).

Inferences from your plots

Females

- **Safe Zone** has the **highest survival chance** for females (overwhelmingly "Survived"; very small zombie proportion).
- **Isolated Island** females also show **high survival**, but with a much smaller total sample.
- **Rural Area** females are roughly **around half survive and half turn into zombies** (much worse than Safe Zone).
- **Contaminated City** females still have **more survivors than zombies**, but survival is lower than Safe Zone.

Males

- For males, **turning into zombies dominates in all locations** (zombie proportion is much higher than survival).
- **Safe Zone** is still the *best among the four* for males (highest survival proportion compared to other locations), but zombie outcomes remain common.
- **Contaminated City** and **Rural Area** show **very poor survival** for males (mostly zombies).
- **Isolated Island** has the **largest male counts**, and the most common outcome is clearly **turning into zombies**.

Overall conclusion

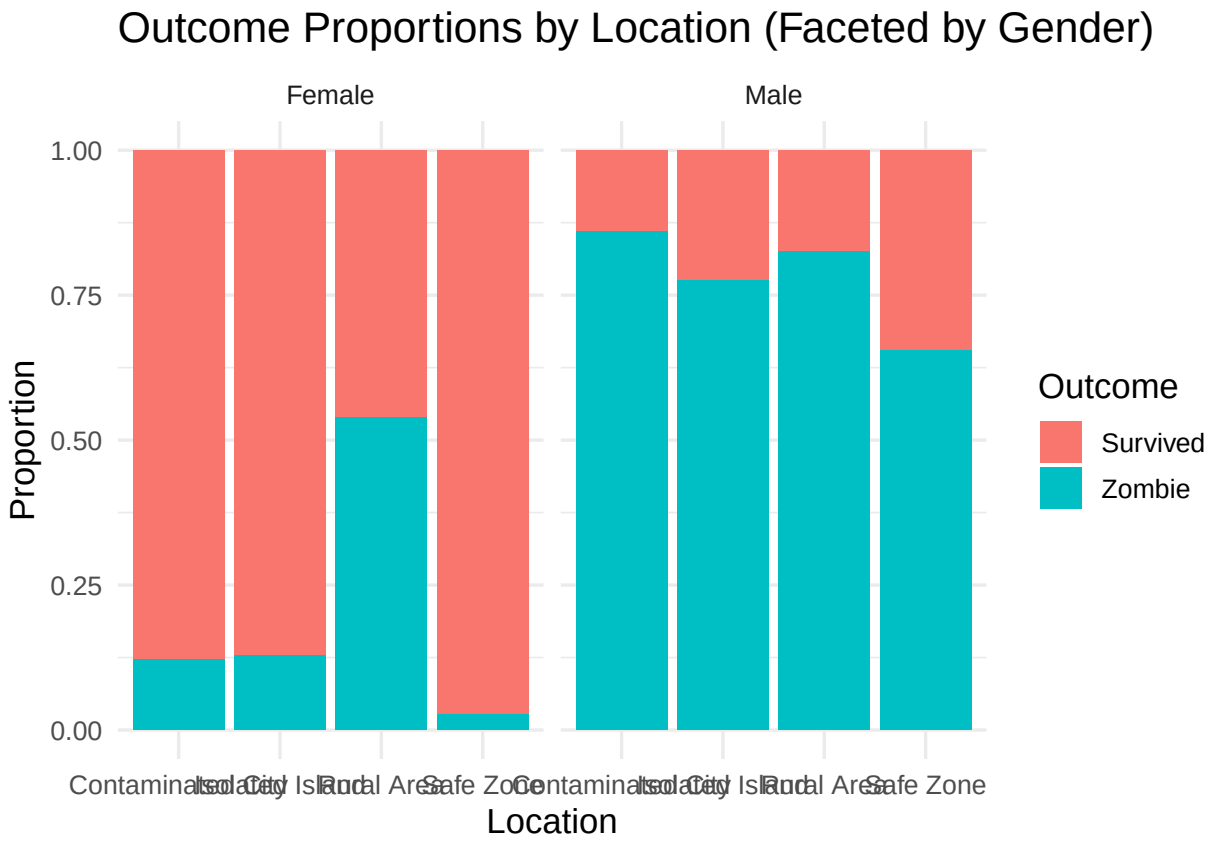
- **Best location for survival overall: Safe Zone** (especially strong for females).
- **Worst survival environment: Rural Area / Contaminated City** (high zombie proportions, especially for males).
- **Most common outcome overall: Turning into zombies**, driven largely by male outcomes and high counts in Rural Area/Isolated Island.

```r
# install.packages(c("tidyr","dplyr","ggplot2","readxl"))

library(tidyr)
library(dplyr)
```

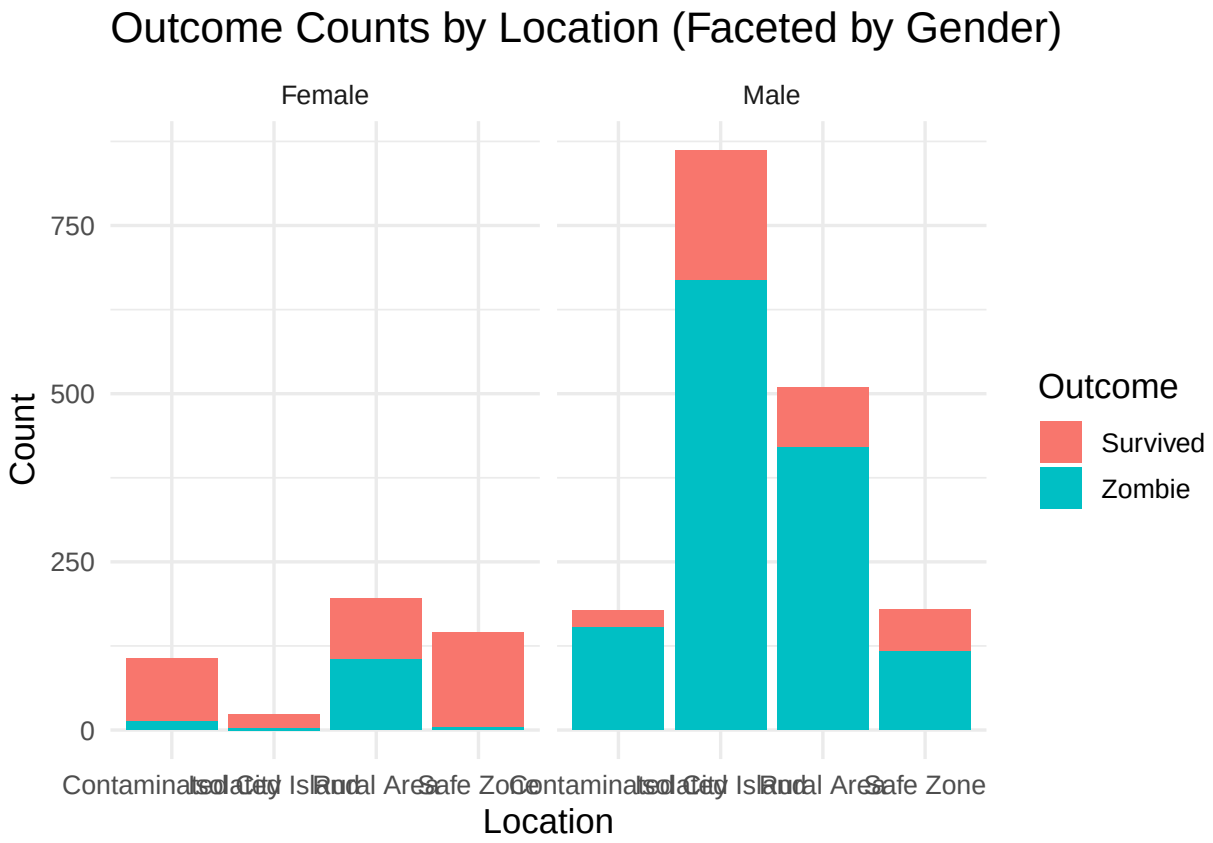
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(readxl)

# =========================
# Q3.1 The Last of Us (given counts)
# =========================
last_of_us <- tribble(
```

```
  ~Location,             ~Gender,  ~Outcome,   ~Count,
  "Safe Zone",           "Male",   "Zombie",   118,
  "Safe Zone",           "Male",   "Survived", 62,
  "Safe Zone",           "Female", "Zombie",   4,
  "Safe Zone",           "Female", "Survived", 141,

  "Contaminated City",   "Male",   "Zombie",   154,
  "Contaminated City",   "Male",   "Survived", 25,
  "Contaminated City",   "Female", "Zombie",   13,
  "Contaminated City",   "Female", "Survived", 93,

  "Rural Area",          "Male",   "Zombie",   422,
  "Rural Area",          "Male",   "Survived", 88,
  "Rural Area",          "Female", "Zombie",   106,
  "Rural Area",          "Female", "Survived", 90,

  "Isolated Island",     "Male",   "Zombie",   670,
  "Isolated Island",     "Male",   "Survived", 192,
  "Isolated Island",     "Female", "Zombie",   3,
  "Isolated Island",     "Female", "Survived", 20
)

# Add totals + survival rate
last_of_us2 <- last_of_us %>%
  group_by(Location, Gender) %>%
  mutate(Total = sum(Count)) %>%
  ungroup() %>%
  mutate(Prop = Count / Total)

# Plot A: Outcome proportions (best for "survival chances")
ggplot(last_of_us2, aes(x = Location, y = Prop, fill = Outcome)) +
  geom_col(position = "fill") +
  facet_wrap(~Gender) +
  labs(title = "Outcome Proportions by Location (Faceted by Gender)",
       x = "Location", y = "Proportion") +
  theme_minimal(base_size = 13)
```




```r
# Plot B: Raw counts (common outcomes, absolute burden)
ggplot(last_of_us, aes(x = Location, y = Count, fill = Outcome)) +
  geom_col(position = "stack") +
  facet_wrap(~Gender) +
  labs(title = "Outcome Counts by Location (Faceted by Gender)",
       x = "Location", y = "Count") +
  theme_minimal(base_size = 13)
```



Question 2:

3.2 Subtask 2: Glass Glimpse Sheet 3 (Glass Glimpse) contains a part of the Glass Classification dataset. Given this dataset, extract the rows pertaining to Glass Type and RI (Refractive Index) and plot visualisations to understand the relationship between them. Mention what inferences you can make from your plot and justify the reasoning behind your choice. The types are numbered 1 to 7 as follows: 1. building-windows-float-processed 2. building-windows-non-float-processed 3. vehicle-windows-float-processed 4. vehicle-windows-non-float-processed (none in this database) 5. containers 6. tableware 7. headlamps

Answer 2:

Visualizations used

1. **Boxplot + points**
2. **Violin + points**

Why these are the right choices

- Glass Type is **categorical (1–7)** and RI is **continuous** → the correct approach is to compare RI **distributions across categories**.
- Boxplot gives **median + spread + outliers**, while violin shows the **shape/density**.

- Adding points is important because it reveals **sample size and clustering** (and prevents hiding structure).

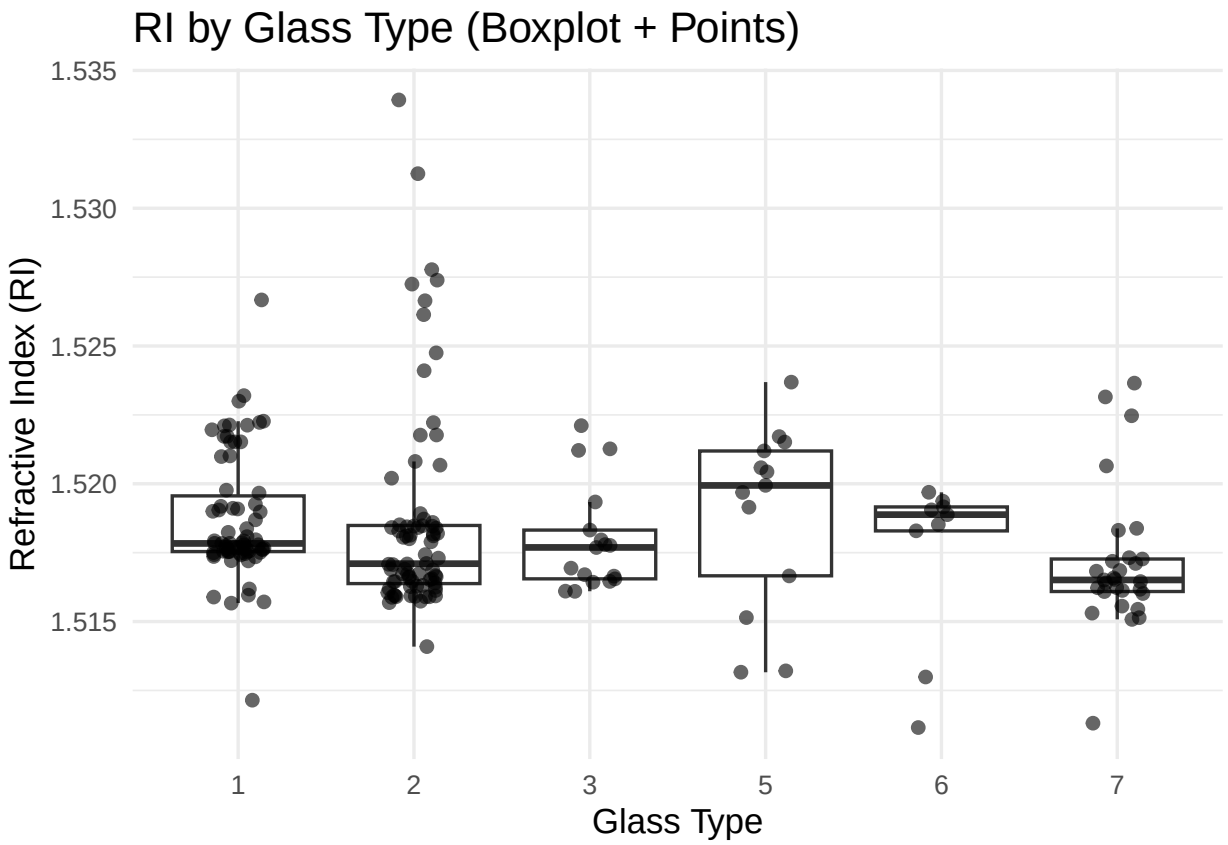
Inferences from your plots

- RI values across types **overlap heavily**, meaning RI alone may not perfectly separate glass types.
- **Type 5 (containers)** appears to have a **slightly higher median RI** than many other types (your boxplot shows this shift upward).
- **Type 6 (tableware)** has a **very tight distribution** (low variability), suggesting RI is quite consistent for that type.
- **Type 2** shows **many samples** and relatively wider spread with some higher outliers (a few RI values extend upward).
- **Type 7** tends to have a **slightly lower central RI** compared to type 5 and shows some spread/outliers.
- **Type 4 is absent**, which matches the statement "none in this database."

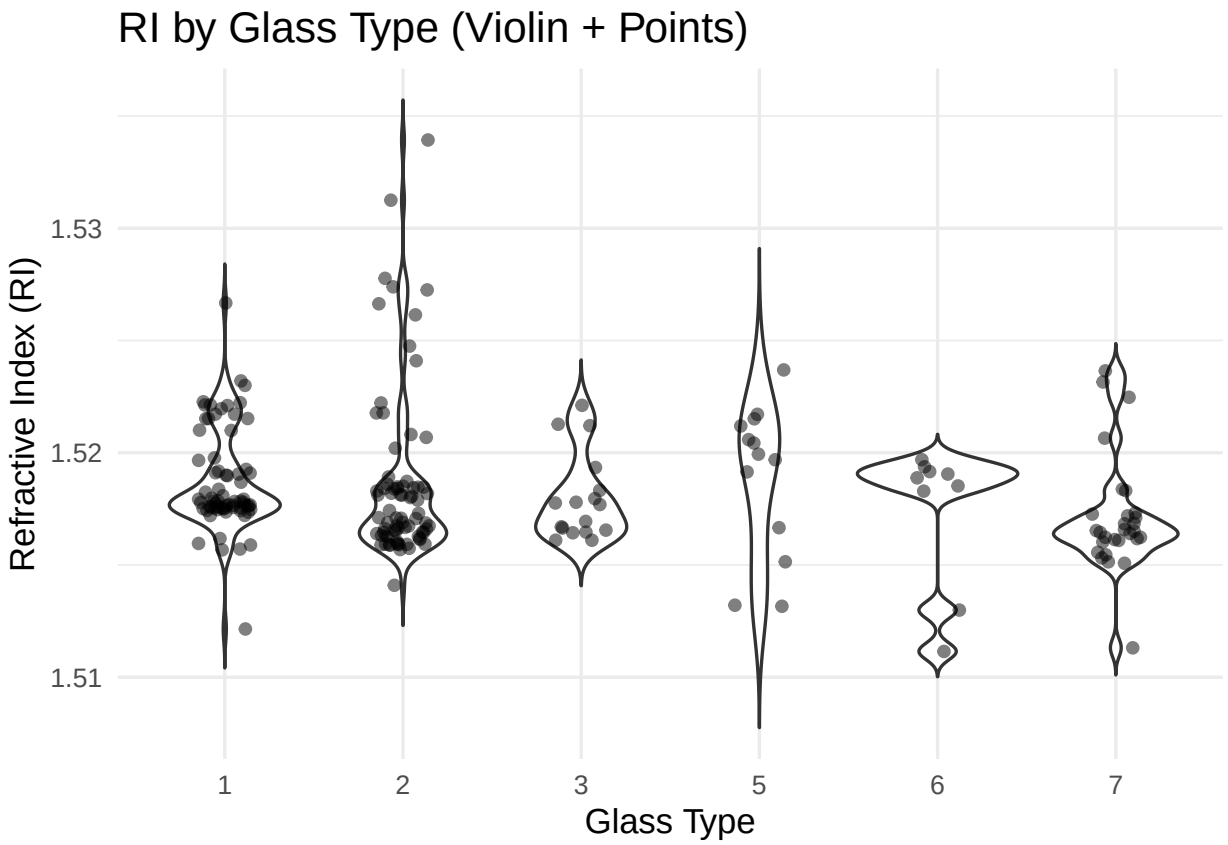
Conclusion

RI shows **some differences in central tendency and variability** between glass types (e.g., type 5 higher, type 6 tight), but because of overlap, RI is best interpreted as a **partial indicator**, not a perfect classifier by itself.

```
# ========================
# Q3.2 Glass Glimpse (Sheet 3): RI vs Glass Type
# ========================
xlsx_path <- "BRSM_Visualisation_Assignment.xlsx"
glass <- read_excel(xlsx_path, sheet = "Glass Glimpse")
glass <- as.data.frame(glass)

# Make flexible column selection (handles slight name differences)
colnames(glass) <- trimws(colnames(glass))

type_col <- grep("^Type$|Glass.*Type|glass.*type", colnames(glass), ignore.case = TRUE, value = TRUE)[1]
ri_col   <- grep("^RI$|Refractive.*Index|refractive.*index", colnames(glass), ignore.case = TRUE, value

glass2 <- glass %>%
  select(Type = all_of(type_col), RI = all_of(ri_col)) %>%
  mutate(Type = as.factor(Type), RI = as.numeric(RI)) %>%
  filter(!is.na(Type), !is.na(RI))

# Plot A: Boxplot + jitter (distribution per type)
ggplot(glass2, aes(x = Type, y = RI)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.15, alpha = 0.6, size = 1.8) +
  labs(title = "RI by Glass Type (Boxplot + Points)",
       x = "Glass Type", y = "Refractive Index (RI)") +
  theme_minimal(base_size = 13)
```




```
# Plot B: Violin + points (shape, multimodality)
ggplot(glass2, aes(x = Type, y = RI)) +
  geom_violin(trim = FALSE) +
  geom_jitter(width = 0.15, alpha = 0.5, size = 1.6) +
  labs(title = "RI by Glass Type (Violin + Points)",
       x = "Glass Type", y = "Refractive Index (RI)") +
  theme_minimal(base_size = 13)
```



Question 3:

3.3 Subtask 3: Night at the Museum Sheet 4 (Museum Visitor) contains a part of the dataset which tracks visitors at the Los Angeles Museums. We focus on five different museums and track visitors from 2014 to 2023. Given this dataset, plot visualisations to understand how the visitor count changes over time. Mention what inferences you can make from your plot and justify the reasoning behind your choice.

Answer 3:

Visualizations used

1. **Multi-line time series plot (all museums together)**
2. **Faceted time series plot (one panel per museum)**

Why these are the right choices

- Visitor count is a **time series** → line plots are the most appropriate for showing **trend, seasonality, and shocks**.
- The combined plot helps compare museums directly, while the faceted plot avoids scale issues and makes each museum's trend easier to see.

Inferences from your plots

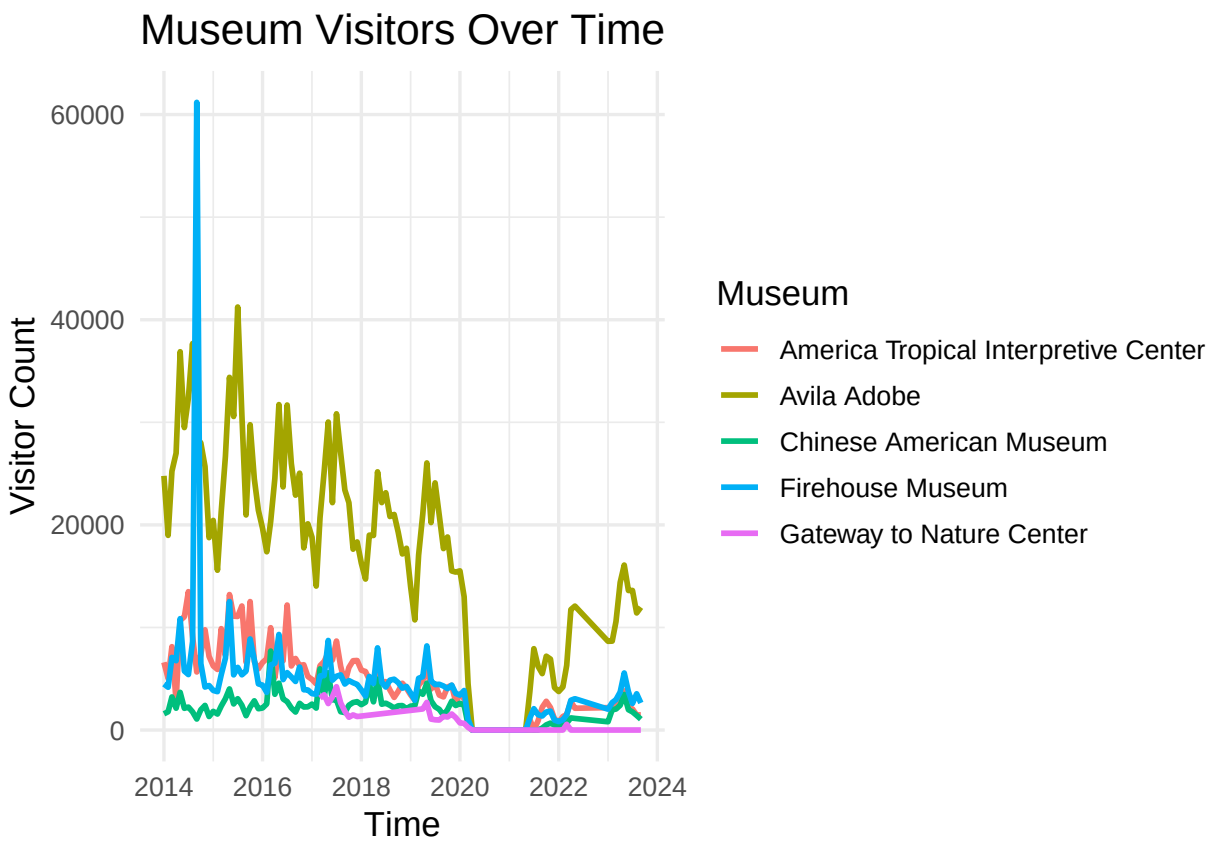
- **Avila Adobe** consistently has the **highest visitor counts** across years (dominant line in the combined plot).
- There is a **sharp drop to near-zero around 2020–2021** across museums, consistent with a major disruption (likely COVID-related closures), and then a **recovery afterward**.
- **Firehouse Museum** has a very large spike early (around 2014) and then stays much lower afterward (suggesting an unusual event/recording spike/outlier).
- **America Tropical Interpretive Center** shows a **gradual decline** toward 2020, then partial recovery post-2021.
- **Chinese American Museum** shows smaller counts overall with variability and recovery after the dip.
- **Gateway to Nature Center** has the **lowest visitor counts** and even reaches near-zero for a period after 2020, indicating prolonged closure or missing/low attendance.

Conclusion

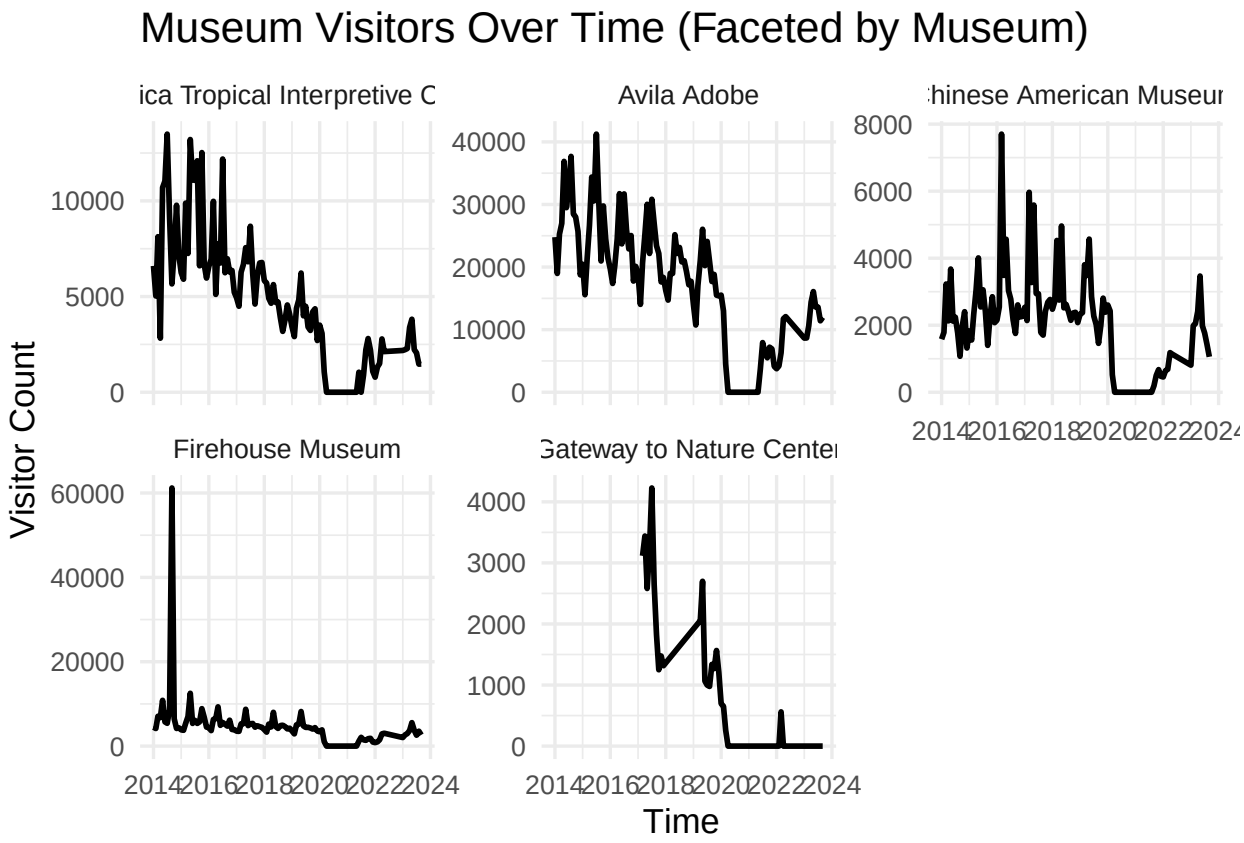
The museums show clear time trends and a strong structural break around 2020. Line plots (especially faceted) effectively reveal both overall comparisons and museum-specific patterns over time.

```r
# ========================
# Q3.3 Night at the Museum
# ========================
library(readxl)
library(tidyr)
library(dplyr)
library(ggplot2)

museum <- read_excel(xlsx_path, sheet = "Museum Visitor")
museum <- as.data.frame(museum)
colnames(museum) <- trimws(colnames(museum))

# Convert Month to Date
museum$Month <- as.Date(paste0("01 ", museum$Month), format = "%d %b %Y")

# Convert from wide to long
museum_long <- museum %>%
  pivot_longer(
    cols = -Month,
    names_to = "Museum",
    values_to = "Visitors"
  ) %>%
  mutate(
    Visitors = as.numeric(Visitors),
    Museum = as.factor(Museum)
  ) %>%
  filter(!is.na(Visitors))

# ---- Plot 1: Overall trends ----
ggplot(museum_long, aes(x = Month, y = Visitors, color = Museum)) +
  geom_line(linewidth = 1) +
  labs(
    title = "Museum Visitors Over Time",
```

```
    x = "Time",
    y = "Visitor Count"
  ) +
  theme_minimal(base_size = 13)
```




```
# ---- Plot 2: Faceted trends ----
ggplot(museum_long, aes(x = Month, y = Visitors)) +
  geom_line(linewidth = 1) +
  facet_wrap(~Museum, scales = "free_y") +
  labs(
    title = "Museum Visitors Over Time (Faceted by Museum)",
    x = "Time",
    y = "Visitor Count"
  ) +
  theme_minimal(base_size = 13)
```


Museum Visitors Over Time (Faceted by Museum)
ica Tropical Interpretive C
Avila Adobe
hinese American Museur
Firehouse Museum
Gateway to Nature Center
Visitor Count
Time
0
5000
10000
20000
30000
40000
60000
1000
2000
3000
4000
6000
8000
2014
2016
2018
2020
2022
2024