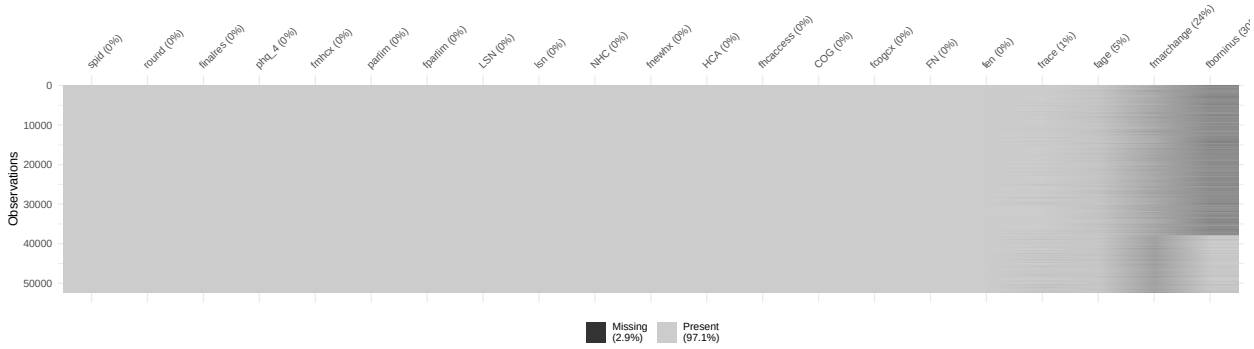
SP EDA

Ruijian Maggie Lin

2025-04-04

```
final_data <- read.csv("SP_rename_final.csv")

# Missingness of raw dataset
vis_miss(final_data, warn_large_data = FALSE)
```



EDA

Our exploratory data analysis (EDA) of the final dataset revealed critical insights through multiple visualizations, including significant trends and underlying data challenges. Key findings—such as severe class imbalances—highlight urgent issues requiring resolution before proceeding to modeling. These visualizations underscore the need for data validation, feature engineering, and balancing strategies to ensure robust model performance.

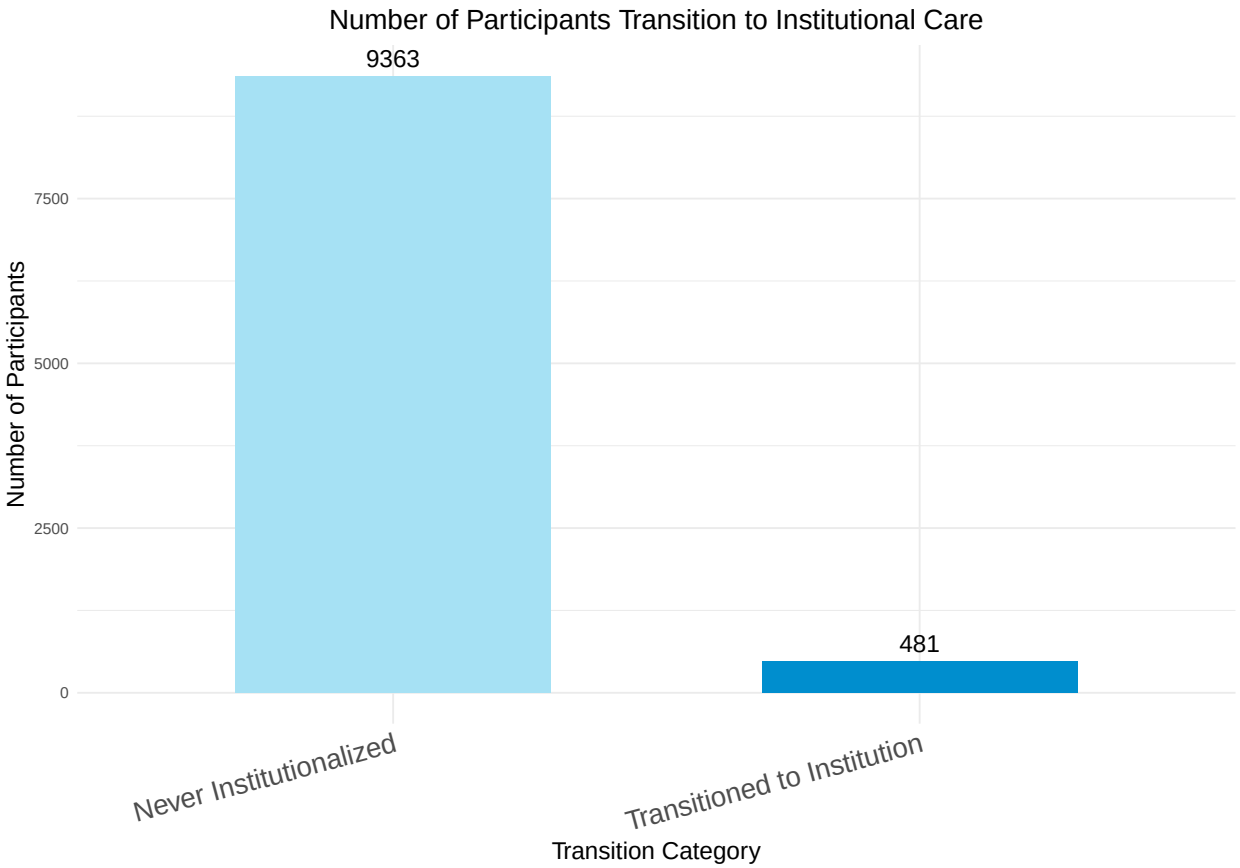
Number of Participants Transitions to Institutional Care

```
# Identify participants who were NEVER institutionalized (finalres = 0 in all rounds)
never_institution <- final_data %>%
  group_by(spid) %>%
  summarize(all_zero = all(finalres == 0)) %>%
  filter(all_zero) %>%
  select(spid)

never_institution_count <- nrow(never_institution)

# Identify participants who TRANSITIONED TO institutional care at the END
```

```r
transition_to_institution <- final_data %>%
  group_by(spid) %>%
  summarize(first_status = first(finalres), last_status = last(finalres)) %>%
  filter(first_status == 0 & last_status == 1) %>%
  select(spid)

transition_to_institution_count <- nrow(transition_to_institution)

# Results
list(
  never_institution_count = never_institution_count,
  transition_to_institution_count = transition_to_institution_count
)
```

```
## $never_institution_count
## [1] 9363
##
## $transition_to_institution_count
## [1] 481
```

```r
# Create a dataframe to store counts
transition_summary <- data.frame(
  Category = c("Never Institutionalized", "Transitioned to Institution"),
  Count = c(never_institution_count, transition_to_institution_count)
)

# Create bar plot
ggplot(transition_summary, aes(x = Category, y = Count, fill = Category)) +
  geom_bar(stat = "identity", width = 0.6) +
  geom_text(aes(label = Count), vjust = -0.5, size = 5) +
  labs(title = "Number of Participants Transition to Institutional Care",
       x = "Transition Category",
       y = "Number of Participants") +
  theme_minimal() +
  theme(legend.position = "none",
       axis.text.x = element_text(angle = 15, hjust = 1, size = 16),
       axis.title.x = element_text(size = 14),
       axis.title.y = element_text(size = 14),
       plot.title = element_text(hjust = 0.5, size = 16)) +
  scale_fill_manual(values = c("#A6E1F4", "#008ECE"))
```



Out of 9844 participants with mental health condition & Complete at least 2 rounds of Survey before institutionalization:

- **Transition to Institutional Care:** 481 participants remained in community in first rounds, and ended institutionalized in the last round of the survey.
- **Never Institutionalized**: 9363 participants were remained in the community in all rounds.

ISSUE: The dataset exhibits severe class imbalance, which would significantly impact model performance if used for predicting transitions to institutional care. Such extreme imbalance would lead models to overfit the majority class and fail to generalize for the minority class.

- Models trained on this data would likely achieve high "accuracy" by always predicting the majority class ("Never Institutionalized"), while failing to identify the minority class ("Transitioned"). For example, a model could appear 95% 'accurate' by simply predicting the majority class every time, which would be unhelpful for identifying the 481 critical cases.

*We might consider a resampling method to address this imbalance dataset in order to perform the models for next step accurately.

Trends in Social Participation Limitation Before Institutionalization (in 9 Rounds)

```r
# Identify participants who completed all 9 rounds
complete_participants <- final_data %>%
  group_by(spid) %>%
  summarize(rounds_completed = n_distinct(round)) %>%
  filter(rounds_completed == 9) %>%
  select(spid)

complete_participants_count <- nrow(complete_participants)
complete_participants_count
```

```
## [1] 2390
```

```r
# Filter the data to include only participants who completed all 9 rounds
final_data_complete <- final_data %>%
  filter(spid %in% complete_participants$spid)

# Identify participants who were NEVER institutionalized (finalres = 0 in all rounds)
never_institution <- final_data_complete %>%
  group_by(spid) %>%
  summarize(all_zero = all(finalres == 0)) %>%
  filter(all_zero) %>%
  select(spid)

never_count <- nrow(never_institution)
never_count
```

```
## [1] 2361
```

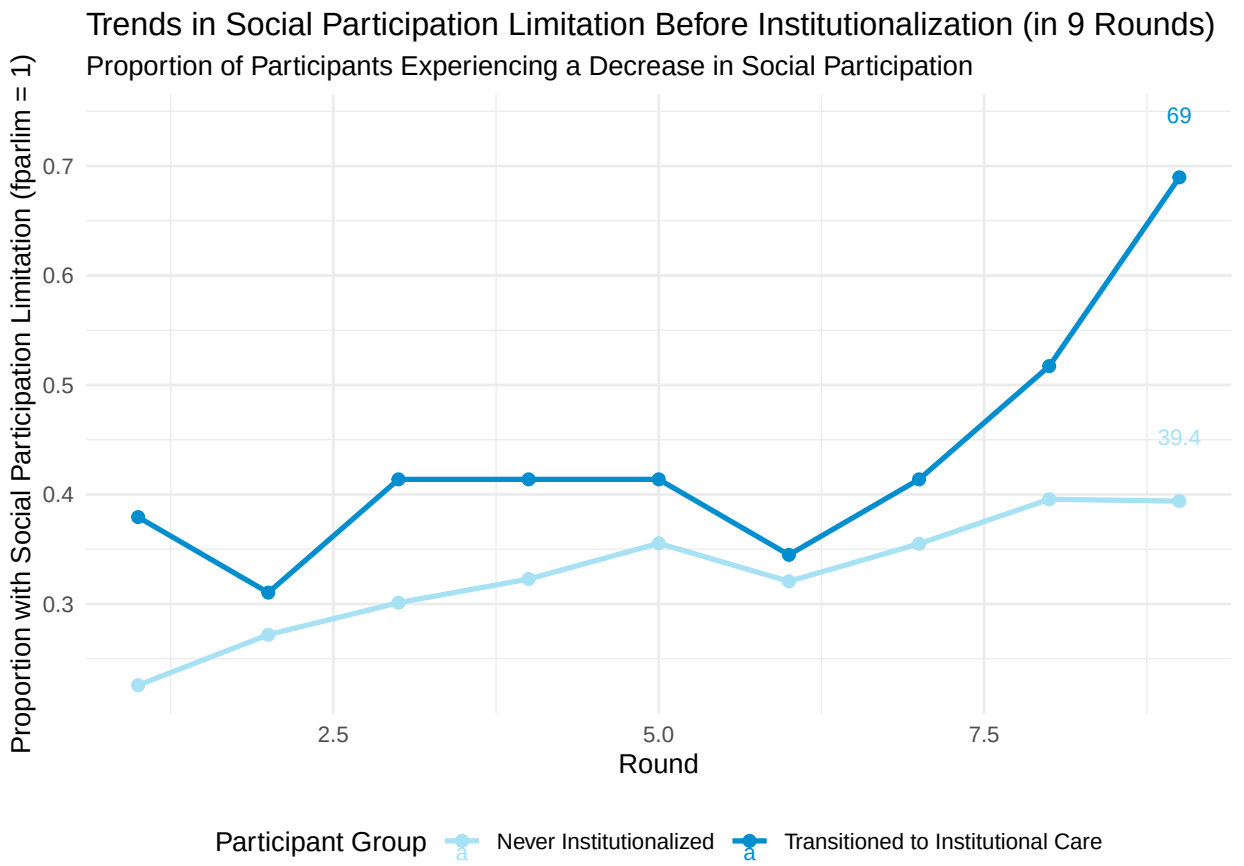
```r
# Identify participants who TRANSITIONED TO institutional care at the END
transition_to_institution <- final_data_complete %>%
  group_by(spid) %>%
  summarize(first_status = first(finalres), last_status = last(finalres)) %>%
  filter(first_status == 0 & last_status == 1) %>%
  select(spid)

transition_count <- nrow(transition_to_institution)
transition_count
```

```
## [1] 29
```

```r
# Identify groups and calculate trends for social participation
trends_data <- final_data_complete %>%
  mutate(group = case_when(
    spid %in% never_institution$spid ~ "Never Institutionalized",
    spid %in% transition_to_institution$spid ~ "Transitioned to Institutional Care",
    TRUE ~ NA_character_ # Exclude other groups
  )) %>%
  filter(!is.na(group)) %>%  # Keep only the identified groups
  group_by(round, group) %>%
  summarize(
    participation_decrease_rate = mean(fparlim, na.rm = TRUE), # Proportion of participants with fparli
```

```
    .groups = 'drop'
  )

# Prepare data for geom_text
text_data <- trends_data %>%
  group_by(group) %>%
  summarize(
    max_y = max(participation_decrease_rate, na.rm = TRUE),
    last_round = last(round),
    participation_decrease_rate = last(participation_decrease_rate),
    .groups = 'drop'
  ) %>%
  mutate(round = last(trends_data$round[trends_data$group == group]))

# Create the trend plot
ggplot(trends_data, aes(x = round, y = participation_decrease_rate, color = group)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  scale_color_manual(values = c("Transitioned to Institutional Care" = "#008ECE",
                                "Never Institutionalized" = "#A6E1F4")) +
  labs(
    title =
      "Trends in Social Participation Limitation Before Institutionalization (in 9 Rounds)",
    subtitle = "Proportion of Participants Experiencing a Decrease in Social Participation",
    x = "Round",
    y = "Proportion with Social Participation Limitation (fparlim = 1)",
    color = "Participant Group"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom") +
  geom_text(data = text_data,
            aes(label = round(participation_decrease_rate * 100, 1), y = max_y + 0.05),
            size = 3, vjust = 0)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



This graph focuses on trends in proportion of participants experienced **decrease in social participation** over time by comparing 2 groups: Those **never institutionalized** and Those who **transitioned to institutional care** in the last round. Both groups include only participants with **complete data for all 9 rounds**, ensuring a fair comparison (There are **2390 participants** who have complete data for all 9 rounds. Out of 2390 participants, **2361** were in the "Never Institutionalized" group, and **29** transitioned to institutional care).

At every round, the "Transitioned" group shows a **higher proportion of participants with decreased social participation** compared to the "Never Institutionalized" group (the line for the "Transitioned" group is consistently above the "Never Institutionalized" group).

The gap between the two groups **widens in later rounds** (e.g., rounds 7–9), this indicates that declining social participation becomes more pronounced as participants approach institutionalization. For example, the gap by **Round 9** shows that **69%** of the "Transitioned" group experienced decreased social participation, but only **39.4%** of the 'Never' group did. This means individuals with decreasing social participation had approximately **1.75 times higher risk of institutionalization**.

Therefore, this graph visually provides strong evidence that **declining social participation predicts institutionalization** in this population.

Social Participation Limitation by New Health Condition Status

```
sp_summary_rename_final <- final_data %>%
  group_by(fnewhx) %>%
```

```
  summarise(sp_count = sum(fparlim == 1, na.rm = TRUE),
            total_count = n(), sp_rate = sp_count /total_count)

sp_summary_rename_final
```

```
## # A tibble: 2 x 4
##   fnewhx sp_count total_count sp_rate
##    <int>    <int>       <int>   <dbl>
## 1      0        7        2959 0.00237
## 2      1    18308       49335 0.371
```

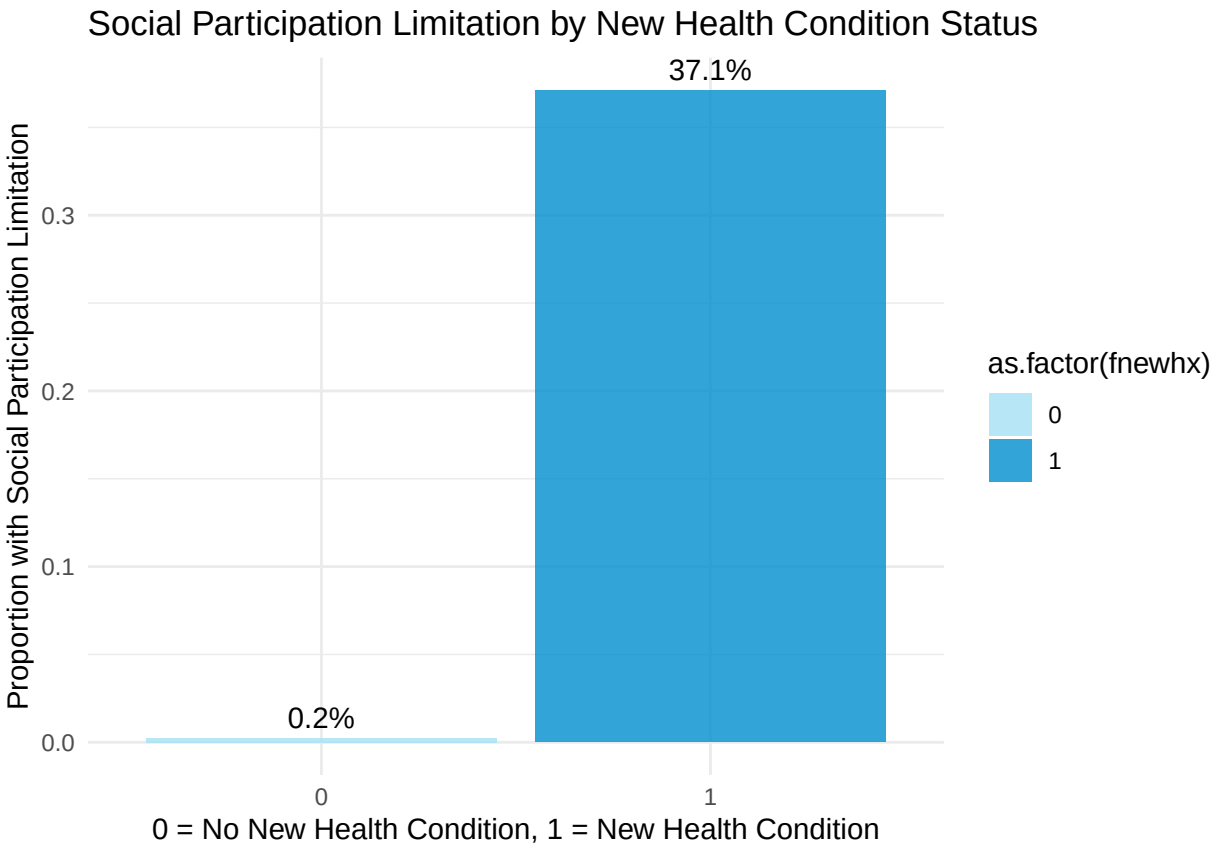
```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
ggplot(sp_summary_rename_final, aes(x = as.factor(fnewhx), y = sp_rate,
                                    fill = as.factor(fnewhx))) +
  geom_bar(stat = "identity", alpha = 0.8) +
  scale_fill_manual(values = c("0" = "#A6E1F4",
                               "1" = "#008ECE")) +
  labs(title =
         "Social Participation Limitation by New Health Condition Status",
       x = "0 = No New Health Condition, 1 = New Health Condition",
       y = "Proportion with Social Participation Limitation") +
  geom_text(aes(label = percent(sp_rate, accuracy = 0.1)), vjust = -0.5) +
  theme_minimal()
```



The graph reveals a **strong association between new health conditions and social participation limitations.**

New Health Condition Present (Group 1): **37.1%** of observations **with** a new health condition reported social participation limitations.

- This aligns with expectations, as new health conditions (e.g., heart attack, high blood pressure, or lung disease) often directly impair social engagement due to physical, emotional, or logistical barriers.

No New Health Condition (Group 0): Only **0.2%** of observations **without** a new health condition reported limitations.

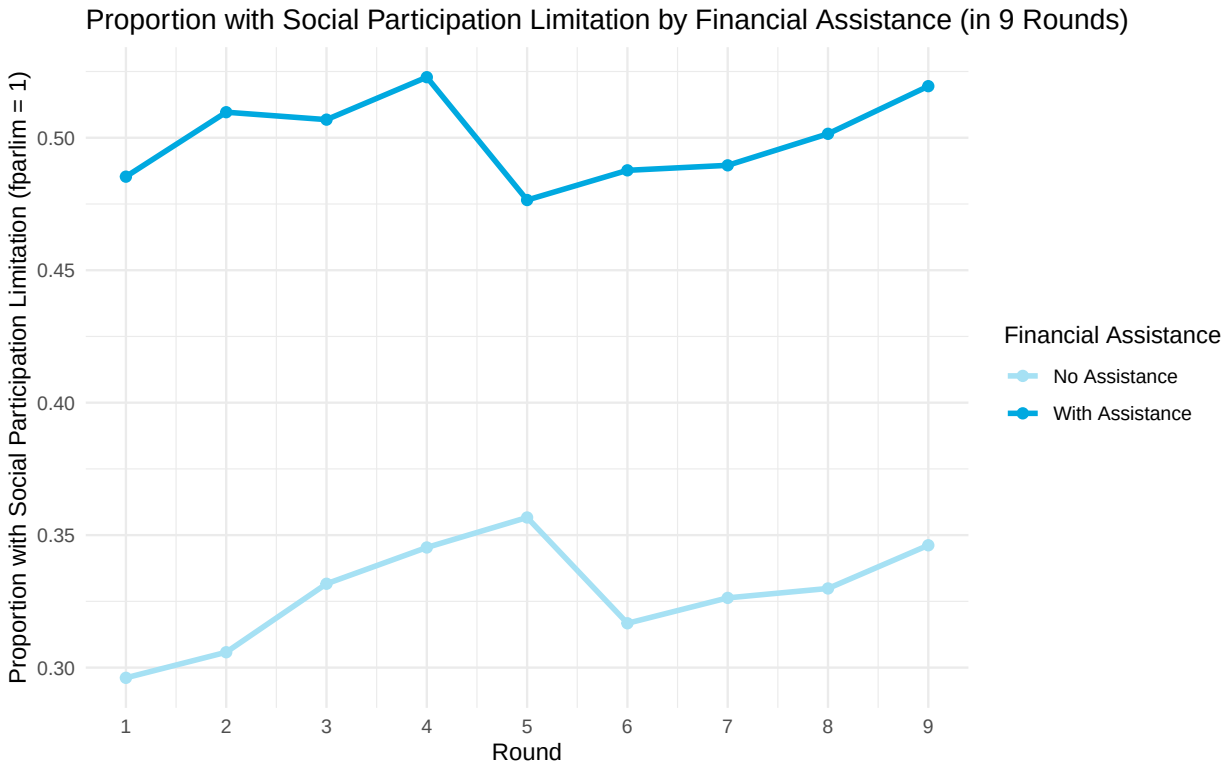
- This suggests that social participation limitations are rare in individuals without recent health changes, reinforcing the idea that health status is a critical driver of such limitations.

Trends in Social Participation Limitation by Financial Assistance (in 9 Rounds)

```
# Group by 'fen' and 'round' and calculate proportion with fparlim = 1
summary_df <- final_data %>%
  group_by(fen, round) %>%
  summarise(prop_fparlim_1 = mean(fparlim == 1, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'fen'. You can override using the `.groups`
## argument.
```

```r
# Plot the proportion with clearer labels
ggplot(summary_df, aes(x = round, y = prop_fparlim_1, color = factor(fen))) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  scale_x_continuous(breaks = 1:9) +
  scale_color_manual(
    values = c("0" = "#A6E1F4", "1" = "#00A9E0"),
    labels = c("0" = "No Assistance", "1" = "With Assistance")
  ) +
  labs(title = "Proportion with Social Participation Limitation by Financial Assistance (in 9 Rounds)",
       x = "Round",
       y = "Proportion with Social Participation Limitation (fparlim = 1)",
       color = "Financial Assistance") +
  theme_minimal()
```

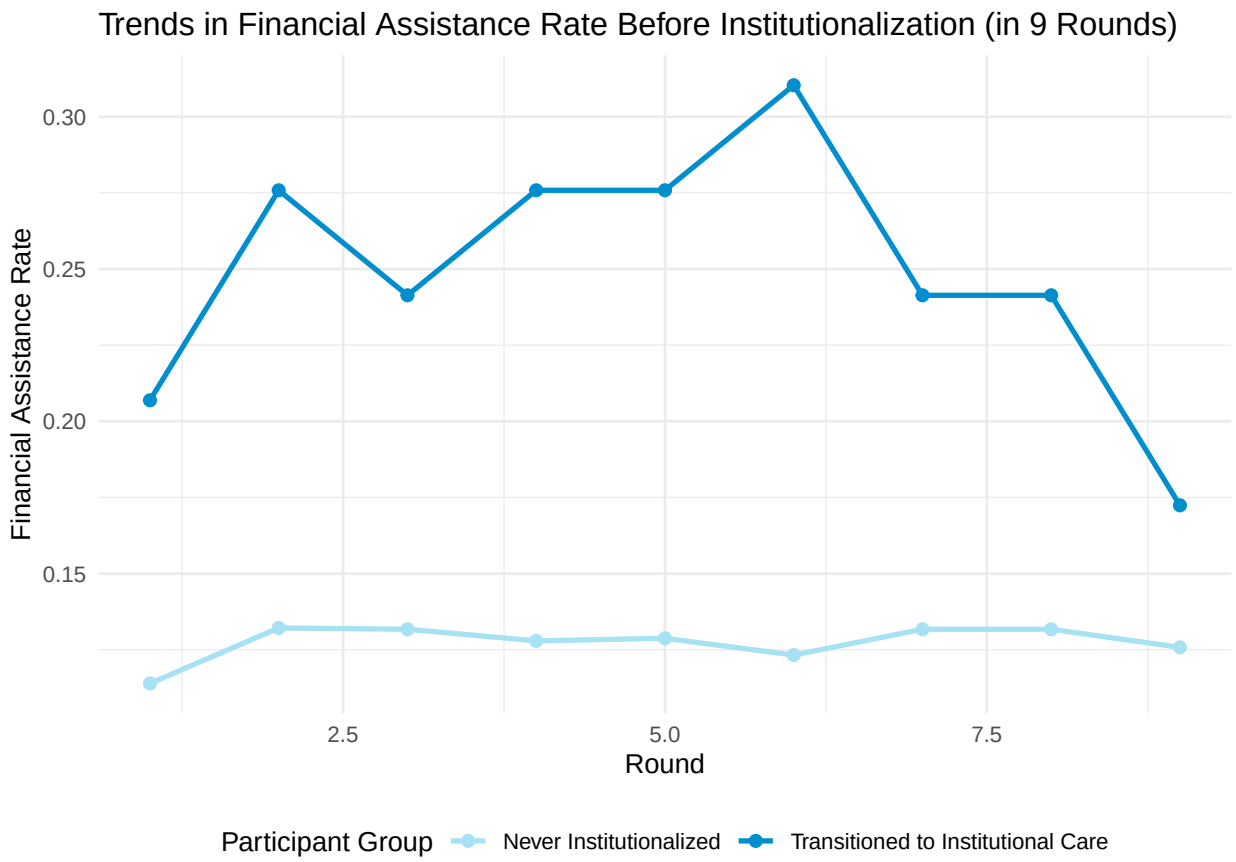


The "With Assistance" group consistently shows a **higher proportion of social participation limitations** compared to the "No Assistance" group across most or all rounds. This suggests that financial assistance recipients report **more social participation limitations** than non-recipients.

- Potential explanations:
  - **Targeted Aid**: Assistance may be directed toward individuals with pre-existing limitations (e.g., disabilities, chronic illnesses), who inherently face greater social barriers.
  - **Unmet Needs**: Financial aid alone might not address root causes of limitations (e.g., accessibility, stigma, or mental health).
  - **Reverse Causality**: Limitations could prompt individuals to seek assistance, rather than assistance causing limitations.

Trends in Financial Assistance Rate Before Institutionalization (in 9 Rounds)

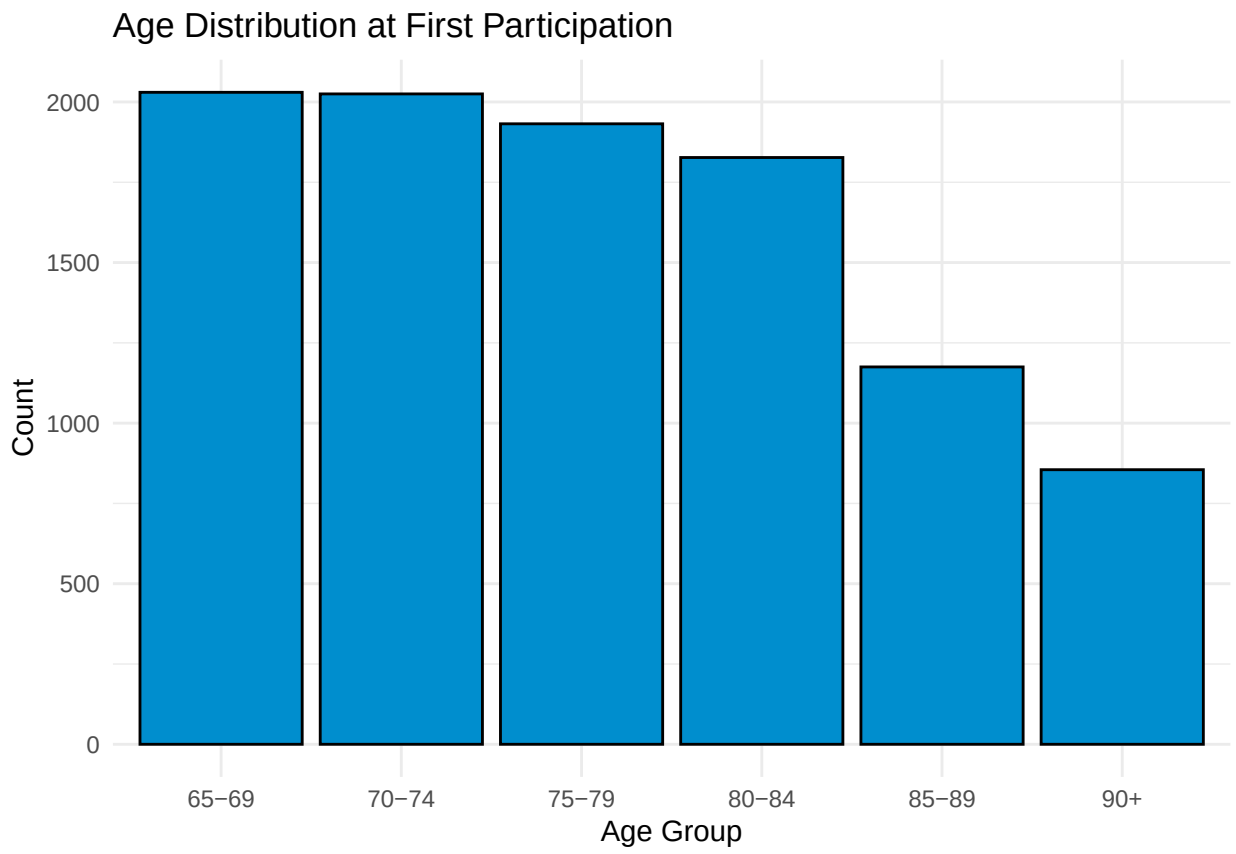
```r
# Identify groups and calculate trends for change in financial need
groups_data <- final_data_complete %>%
  mutate(group = case_when(
    spid %in% never_institution$spid ~ "Never Institutionalized",
    spid %in% transition_to_institution$spid ~ "Transitioned to Institutional Care",
    TRUE ~ NA_character_  # Exclude other groups
  )) %>%
  filter(!is.na(group)) %>%
  group_by(round, group) %>%
  summarize(financial_need_rate = mean(fen, na.rm = TRUE), .groups = 'drop')

# Prepare data for geom_text
text_data <- groups_data %>%
  group_by(group) %>%
  summarize(max_y = max(financial_need_rate, na.rm = TRUE),
            last_round = last(round),
            financial_need_rate = last(financial_need_rate),
            .groups = 'drop') %>%
  mutate(round = last(groups_data$round[groups_data$group == group]))

# Create the trend plot
ggplot(groups_data, aes(x = round, y = financial_need_rate, color = group)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  scale_color_manual(values = c("Transitioned to Institutional Care" = "#008ECE",
                                "Never Institutionalized" = "#A6E1F4")) +
  labs(title = "Trends in Financial Assistance Rate Before Institutionalization (in 9 Rounds)",
       x = "Round",
       y = "Financial Assistance Rate",
       color = "Participant Group") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Based on the chart above, "transitioned" group shows a consistently higher financial assistance rate over the 9 rounds compared to the "never" group. "Never" group maintains a stable and lower financial assistance rate, indicating fewer financial needs or effective financial stability that may contribute to avoiding institutional care. As a result, it suggests that individuals who eventually transitioned to institutional care required increasing financial support prior to institutionalization, possibly due to escalating needs (e.g., medical costs, reduced independence).

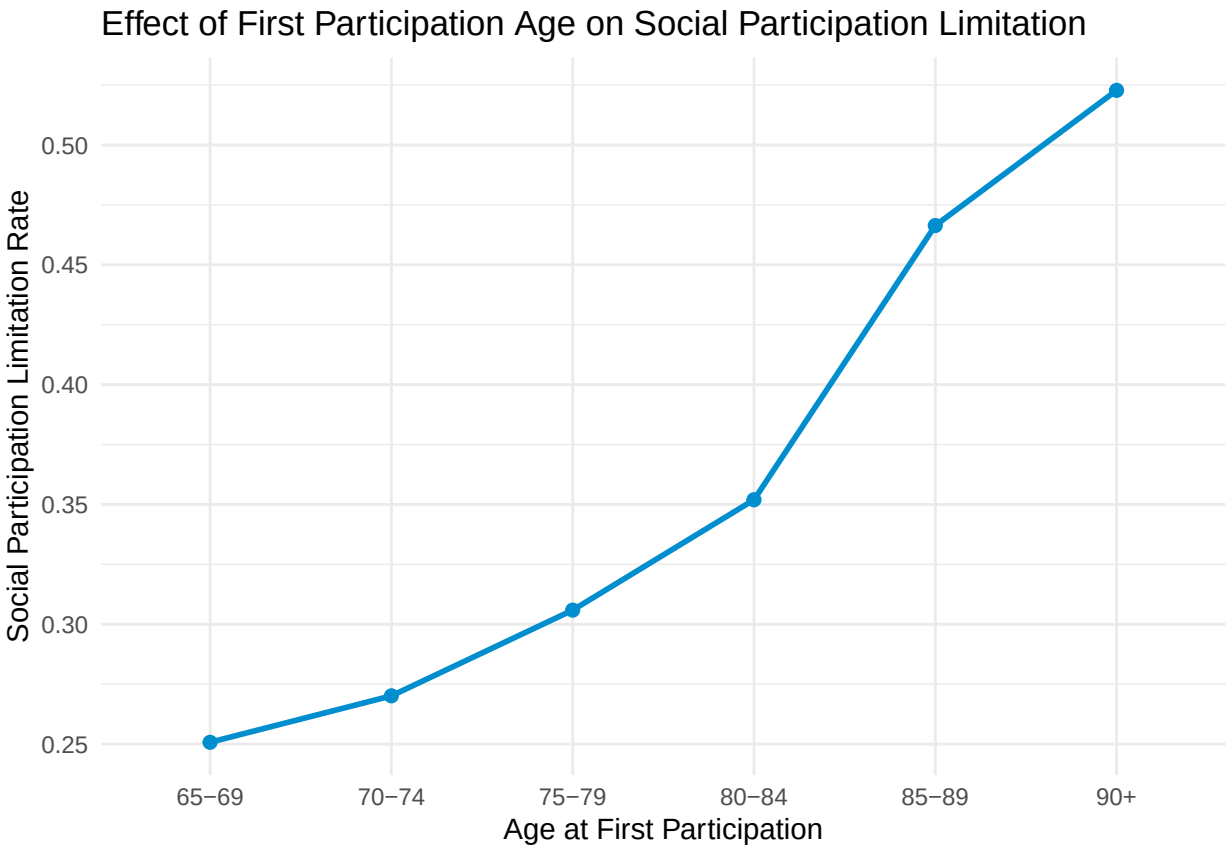
Distribution of Ages at First Round

```
# Extract age at first participation
first_age_data <- final_data %>%
  group_by(spid) %>%
  slice_min(round) %>%  # Extract the earliest round
  ungroup() %>%
  select(spid, fage)  # Select ID and age columns

# Plot the age distribution
ggplot(first_age_data, aes(x = fage)) +
  geom_bar(fill = "#008ECE", color = "black") +
  labs(title = "Age Distribution at First Participation",
       x = "Age Group",
       y = "Count") +
  theme_minimal()
```



This bar chart shows the distribution of participants' ages at their first survey participation. The majority of participants were aged 65-79, with fewer individuals in older age groups. Notably, participation significantly declines in the 85-89 and 90+ age groups, which may reflect reduced engagement or eligibility in later life stages.

```
age_participation_data <- final_data %>%
  group_by(spid) %>%
  slice_min(round) %>%  # Extract the earliest round for each participant
  ungroup() %>%
  group_by(fage) %>%
  summarize(participation_limitation_rate = mean(fparlim, na.rm = TRUE), .groups = 'drop') %>%
  drop_na()  # Remove NA values to ensure a continuous line

# Ensure fage is treated as an ordered factor for proper line connection
age_participation_data <- age_participation_data %>%
  mutate(fage = factor(fage, levels = c("65-69", "70-74", "75-79", "80-84", "85-89", "90+"), ordered = T

# Plot the trend of social participation limitation by first participation age
ggplot(age_participation_data, aes(x = fage, y = participation_limitation_rate, group = 1)) +
  geom_line(color = "#008ECE", size = 1) +
  geom_point(color = "#008ECE", size = 2) +
  labs(title = "Effect of First Participation Age on Social Participation Limitation",
       x = "Age at First Participation",
       y = "Social Participation Limitation Rate") +
  theme_minimal()
```



This chart shows that social participation limitations increase with age at first participation. Individuals who joined the survey at younger ages, such as 65-69, had a lower limitation rate (25%), while those who joined at 90+ experienced a significantly higher rate (over 50%). This trend suggests that older age is associated with a greater risk of social participation limitations, emphasizing the importance of early engagement strategies to maintain social involvement.

Imbalance Dataset

```
set.seed(42)

# Step 1: Train-test split at the participant level
participants <- final_data %>% distinct(spid)
train_index <- createDataPartition(participants$spid, p = 0.8, list = FALSE)  # 80% training
train_participants <- participants$spid[train_index]
test_participants <- participants$spid[-train_index]

train_data <- final_data %>% filter(spid %in% train_participants)
test_data <- final_data %>% filter(spid %in% test_participants)

# Step 2: Identify "Never Institutionalized" and "Transitioned" in training set
train_never_institution <- train_data %>%
  group_by(spid) %>%
  summarise(all_zero = all(finalres == 0)) %>%
  filter(all_zero) %>%
```

```
  select(spid)

train_transition_to_institution <- train_data %>%
  group_by(spid) %>%
  summarise(first_status = first(finalres), last_status = last(finalres)) %>%
  filter(first_status == 0 & last_status == 1) %>%
  select(spid)

# Check the number of participants in each group
cat("Number of Never Institutionalized participants in training set:", nrow(train_never_institution), "
```

```
## Number of Never Institutionalized participants in training set: 7490
```

```
cat("Number of Transitioned participants in training set:", nrow(train_transition_to_institution), "\n"
```

```
## Number of Transitioned participants in training set: 386
```

```
# Check unique participants in training data
unique_participants <- train_data %>%
  filter(spid %in% train_never_institution$spid) %>%
  distinct(spid)

cat("Unique Never Institutionalized participants before clustering:", nrow(unique_participants), "\n")
```

```
## Unique Never Institutionalized participants before clustering: 7490
```

```
# Step 3: Perform Cluster-Based Undersampling on Training Set Only
train_never_institution_data <- train_data %>%
  filter(spid %in% train_never_institution$spid)

# Compute summary stats for clustering
train_never_summary <- train_never_institution_data %>%
  group_by(spid) %>%
  summarise(avg_finalres = mean(finalres), num_observations = n(), .groups = "drop")

# Check the unique participants available for clustering
n_unique <- nrow(train_never_summary)
cat("Number of unique Never Institutionalized participants:", n_unique, "\n")
```

```
## Number of unique Never Institutionalized participants: 7490
```

```
# K-means clustering
k <- min(nrow(train_transition_to_institution), n_unique)
clusters <- kmeans(train_never_summary, centers = k)

# Create a data frame with cluster assignments
train_never_summary <- train_never_summary %>%
  mutate(cluster = clusters$cluster)

# Sample one participant from each cluster
```

```r
selected_majority <- train_never_summary %>%
  group_by(cluster) %>%
  sample_n(1) %>%
  ungroup()

# Step 4: Combine selected participants with Transitioned group
final_balanced_training_set <- train_data %>%
  filter(spid %in% selected_majority$spid) %>%
  bind_rows(train_data %>% filter(spid %in% train_transition_to_institution$spid))

# Check the number of participants in each group
num_never_institution <- final_balanced_training_set %>%
  filter(spid %in% train_never_institution$spid) %>%
  summarise(n = n_distinct(spid)) %>%
  pull(n)

num_transitioned <- final_balanced_training_set %>%
  filter(spid %in% train_transition_to_institution$spid) %>%
  summarise(n = n_distinct(spid)) %>%
  pull(n)

# Print results to verify balance
cat("Number of Never Institutionalized participants in balanced training set:", num_never_institution,
```

```
## Number of Never Institutionalized participants in balanced training set: 386
```

```r
cat("Number of Transitioned participants in balanced training set:", num_transitioned, "\n")
```

```
## Number of Transitioned participants in balanced training set: 386
```

```r
# Step 1: Check the number of participants in each group in the testing set
test_never_institution <- test_data %>%
  group_by(spid) %>%
  summarise(all_zero = all(finalres == 0)) %>%
  filter(all_zero) %>%
  select(spid)

test_transition_to_institution <- test_data %>%
  group_by(spid) %>%
  summarise(first_status = first(finalres), last_status = last(finalres)) %>%
  filter(first_status == 0 & last_status == 1) %>%
  select(spid)

# Print results to verify group distribution in the testing set
num_test_never_institution <- nrow(test_never_institution)
num_test_transitioned <- nrow(test_transition_to_institution)

cat("Number of Never Institutionalized participants in testing set:", num_test_never_institution, "\n")
```

```
## Number of Never Institutionalized participants in testing set: 1873
```

```r
cat("Number of Transitioned participants in testing set:", num_test_transitioned, "\n")
```

```
## Number of Transitioned participants in testing set: 95
```

```r
# Step 2: Compare with the balanced training set
cat("Number of Never Institutionalized participants in balanced training set:", num_never_institution,
```

```
## Number of Never Institutionalized participants in balanced training set: 386
```

```r
cat("Number of Transitioned participants in balanced training set:", num_transitioned, "\n")
```

```
## Number of Transitioned participants in balanced training set: 386
```

```r
## Data leakage prevention: Check for overlapping participants
overlap <- intersect(train_data$spid, test_data$spid)
if (length(overlap) == 0) {
  cat("No overlapping participants between train and test sets.")
} else {
  cat("Data leakage detected! Overlapping SPIDs:", overlap)
}
```

```
## No overlapping participants between train and test sets.
```

Problem: Our dataset is "imbalanced": one group (participants who transitioned to institutionalization) is much smaller than the other (participants never institutionalized). This imbalance can bias the model to prioritize the majority class, reducing its ability to predict transitions accurately.

Solution [1]: We used **cluster-based undersampling** to balance the training data while preserving critical patterns:

1. **Participant-Level Split:**
   - Separated training/test data by participant (`spid`) to avoid data leakage and ensure realistic evaluation.
2. **Targeted Balancing:**
   - Identified two key groups: "Never Institutionalized" (majority) and "Transitioned" (minority).
   - Applied **clustering** to group similar "Never Institutionalized" participants (e.g., by average outcome, observation count).
   - Sampled one participant per cluster to retain diversity while reducing the majority's size to match the minority.
3. **Final Dataset:**
   - Combined the undersampled majority with all minority participants to create a balanced training set.

Why This Works:

- Maintains the temporal structure of longitudinal data.

- Avoids bias by ensuring the model learns from both groups equally.
- Clustering preserves meaningful variation in the majority class.

Outcome:
A robust model that fairly represents both groups, improving predictions for transitions while generalizing to real-world scenarios.

References

[1] Practical Guide to Deal with Imbalanced Classification Problems in R. https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/