Exploratory Analysis on Missing Data

Guillem Olivart Garrofé

2025-12-30

Contents

1. Libraries 1
2. Data Preparation 1
3. Exploratory Analysis on Missing Data 3
4. Missingness per Covariates 7
5. Mean Profiles by Dropout Time 15
6. Dropout Mechanism Analysis 19
7. Baseline Comparison: Completers vs Dropouts 25
8. Additional Visualizations 26

1. Libraries

```
library(haven)      # Dades SAS
library(dplyr)      # Manipulació Dades
library(tidyr)      # Transformacions (pivot_longer)
library(ggplot2)    # Gràfics
library(naniar)     # Missing Data
library(VIM)        # Visualització Patrons
library(nlme)       # LMM
library(lme4)       # GLMM
library(geepack)    # GEE i WGEE
library(mice)       # Sensitivity Analysis
library(corrplot)   # Matriu Correlació
```

2. Data Preparation

```
# DATASET
# Assegura't que la ruta és correcta
alz <- read_sas("C:/Users/win11/Documents/Guillem/Erasmus/Assignatures/Longitudinal Data Analysis/Proje

# Factor variables
alz$trial <- as.factor(alz$trial)
```

```r
alz$sex <- as.factor(alz$sex)
alz$edu <- as.factor(alz$edu)
alz$job <- as.factor(alz$job)
alz$wzc <- as.factor(alz$wzc)
alz$adl <- as.factor(alz$adl)
alz$adl_num <- as.numeric(alz$adl)
alz$n_obs_data <- rowSums(!is.na(alz[, c(18:24)]))

# Bins
alz$cdrsb_bin0 <- ifelse(alz$cdrsb0 > 10, 1, 0)
alz$cdrsb_bin1 <- ifelse(alz$cdrsb1 > 10, 1, 0)
alz$cdrsb_bin2 <- ifelse(alz$cdrsb2 > 10, 1, 0)
alz$cdrsb_bin3 <- ifelse(alz$cdrsb3 > 10, 1, 0)
alz$cdrsb_bin4 <- ifelse(alz$cdrsb4 > 10, 1, 0)
alz$cdrsb_bin5 <- ifelse(alz$cdrsb5 > 10, 1, 0)
alz$cdrsb_bin6 <- ifelse(alz$cdrsb6 > 10, 1, 0)

# Baseline
alz$ab_base <- alz$abpet0
alz$tau_base <- alz$taupet0
alz$cdrsb_base <- alz$cdrsb0
alz$bprs_base <- alz$bprs0
```

Longitudinal Dataset

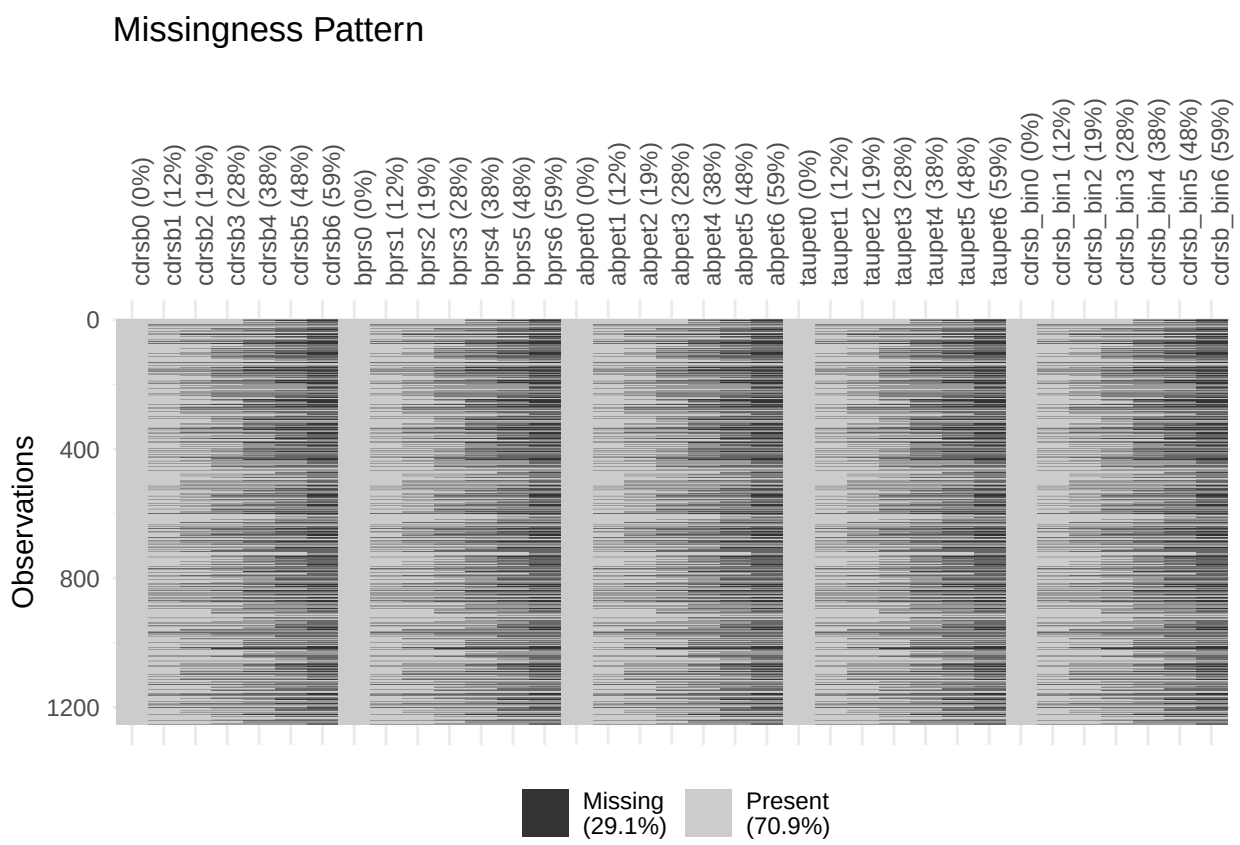
```r
alz_long <- alz %>%
  pivot_longer(
    cols = matches(".*\\d+$"),
    names_to = c(".value", "time"),
    names_pattern = "(.*)(\\d+)$"
  ) %>%
  mutate(
    time = as.numeric(time),
    id = as.factor(patid)
  ) %>%
  arrange(id, time)

head(alz_long)
```

```
## # A tibble: 6 x 23
##   patid trial sex     age edu     bmi inkomen job   adl   wzc   adl_num
##   <dbl> <fct> <fct> <dbl> <fct> <dbl>   <dbl> <fct> <fct> <fct>   <dbl>
## 1 10001 1     0        72 3      26.4    1900 0     6     0           7
## 2 10001 1     0        72 3      26.4    1900 0     6     0           7
## 3 10001 1     0        72 3      26.4    1900 0     6     0           7
## 4 10001 1     0        72 3      26.4    1900 0     6     0           7
## 5 10001 1     0        72 3      26.4    1900 0     6     0           7
## 6 10001 1     0        72 3      26.4    1900 0     6     0           7
## # i 12 more variables: n_obs_data <dbl>, ab_base <dbl>, tau_base <dbl>,
## #   cdrsb_base <dbl>, bprs_base <dbl>, time <dbl>, cdrsb <dbl>, bprs <dbl>,
## #   abpet <dbl>, taupet <dbl>, cdrsb_bin <dbl>, id <fct>
```

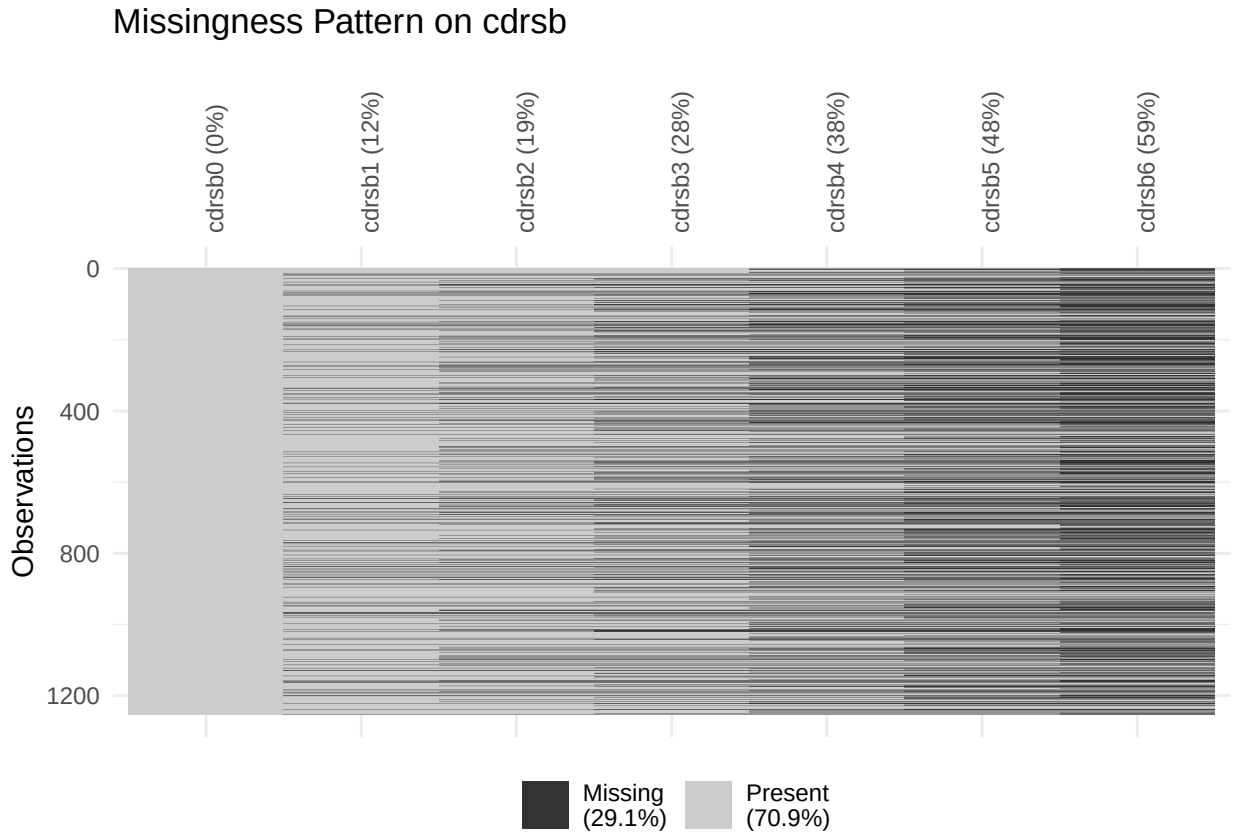
3. Exploratory Analysis on Missing Data

General Behaviour

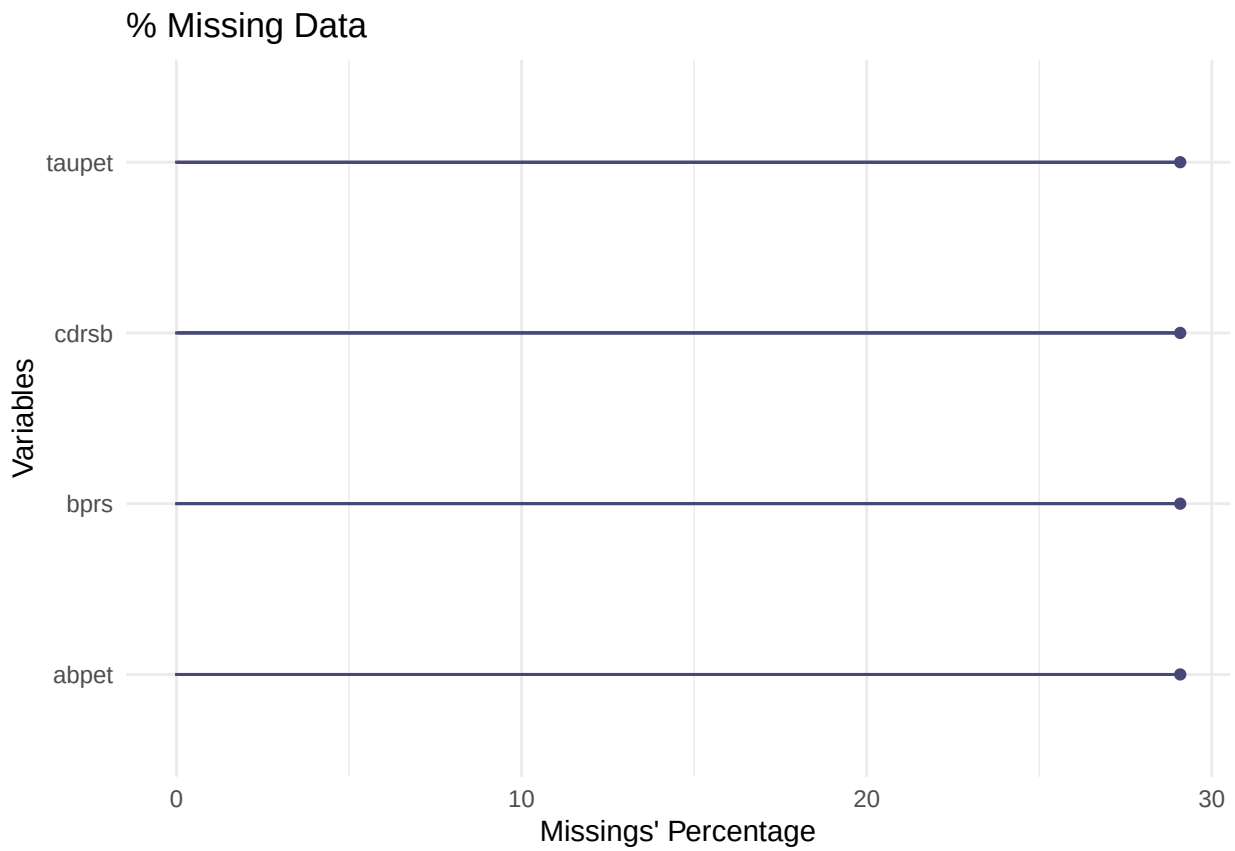
```
# General pattern
alz %>%
  select(matches("\\d+$")) %>%
  vis_miss() +
  labs(title = "Missingness Pattern") +
  theme(axis.text.x = element_text(angle = 90))
```




```
# Specific to CDRSB
alz %>%
  select(matches("cdrsb\\d+$")) %>%
  vis_miss() +
  labs(title = "Missingness Pattern on cdrsb") +
  theme(axis.text.x = element_text(angle = 90))
```




```
# Percentages
alz_long %>%
  select(cdrsb, bprs, abpet, taupet) %>%
  gg_miss_var(show_pct = TRUE) +
  labs(title = "% Missing Data",
       y = "Missings' Percentage")
```



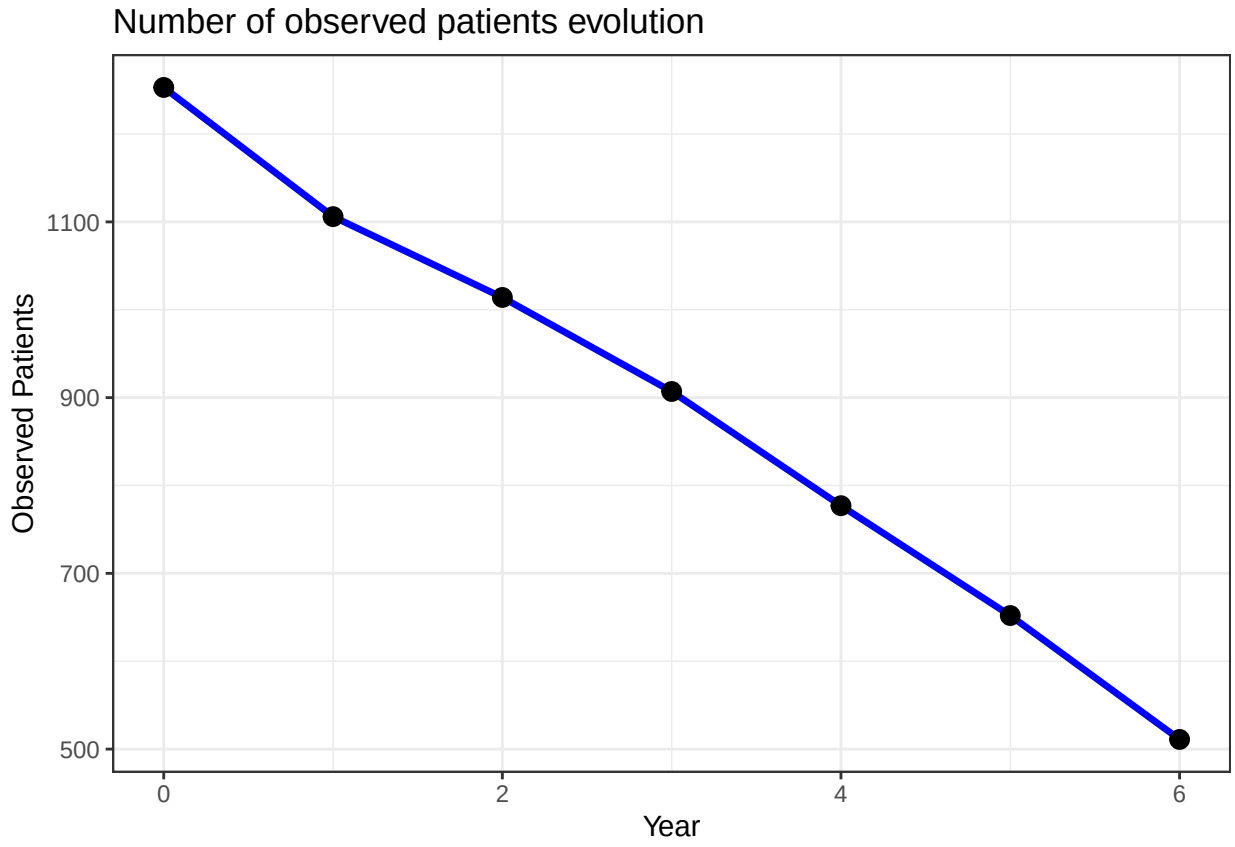
The missingness pattern is **equal** in every longitudinal variable.

Dropout Analysis

```
dropout_table <- alz_long %>%
  group_by(time) %>%
  summarise(
    N_Total = n(),
    Observed_CDRSB = sum(!is.na(cdrsb)),
    Missing_CDRSB = sum(is.na(cdrsb)),
    Pct_Missing = round(mean(is.na(cdrsb)) * 100, 2)
  )

print(dropout_table)
```

```
## # A tibble: 7 x 5
##    time N_Total Observed_CDRSB Missing_CDRSB Pct_Missing
##   <dbl>   <int>          <int>         <int>       <dbl>
## 1     0    1253           1253             0         0
## 2     1    1253           1106           147        11.7
## 3     2    1253           1014           239        19.1
## 4     3    1253            907           346        27.6
## 5     4    1253            777           476        38.0
## 6     5    1253            652           601        48.0
## 7     6    1253            511           742        59.2
```

```r
ggplot(dropout_table, aes(x = time, y = Observed_CDRSB)) +
  geom_line(color = "blue", size = 1.2) +
  geom_point(size = 3) +
  labs(title = "Number of observed patients evolution",
       x = "Year", y = "Observed Patients") +
  theme_bw()
```



Monotoneness Check

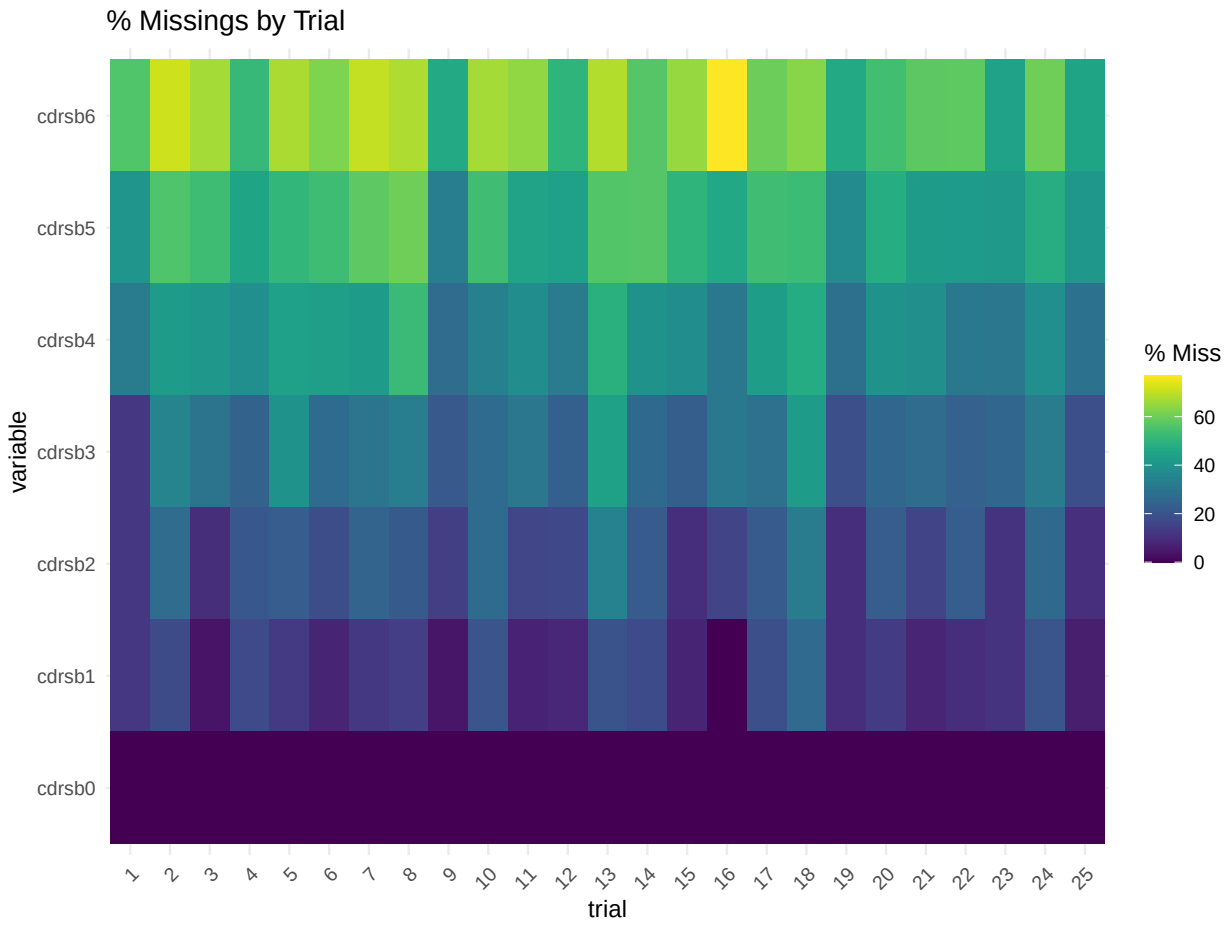
```r
check_intermittent <- alz_long %>%
  select(id, time, cdrsb) %>%
  arrange(id, time) %>%
  group_by(id) %>%
  summarise(
    is_missing = list(is.na(cdrsb)),
    pattern = paste(as.integer(!is.na(cdrsb)), collapse = "")
  )
table(check_intermittent$pattern)
```

```
##
## 1000000 1100000 1110000 1111000 1111100 1111110 1111111
##     147      92     107     130     125     141     511
```

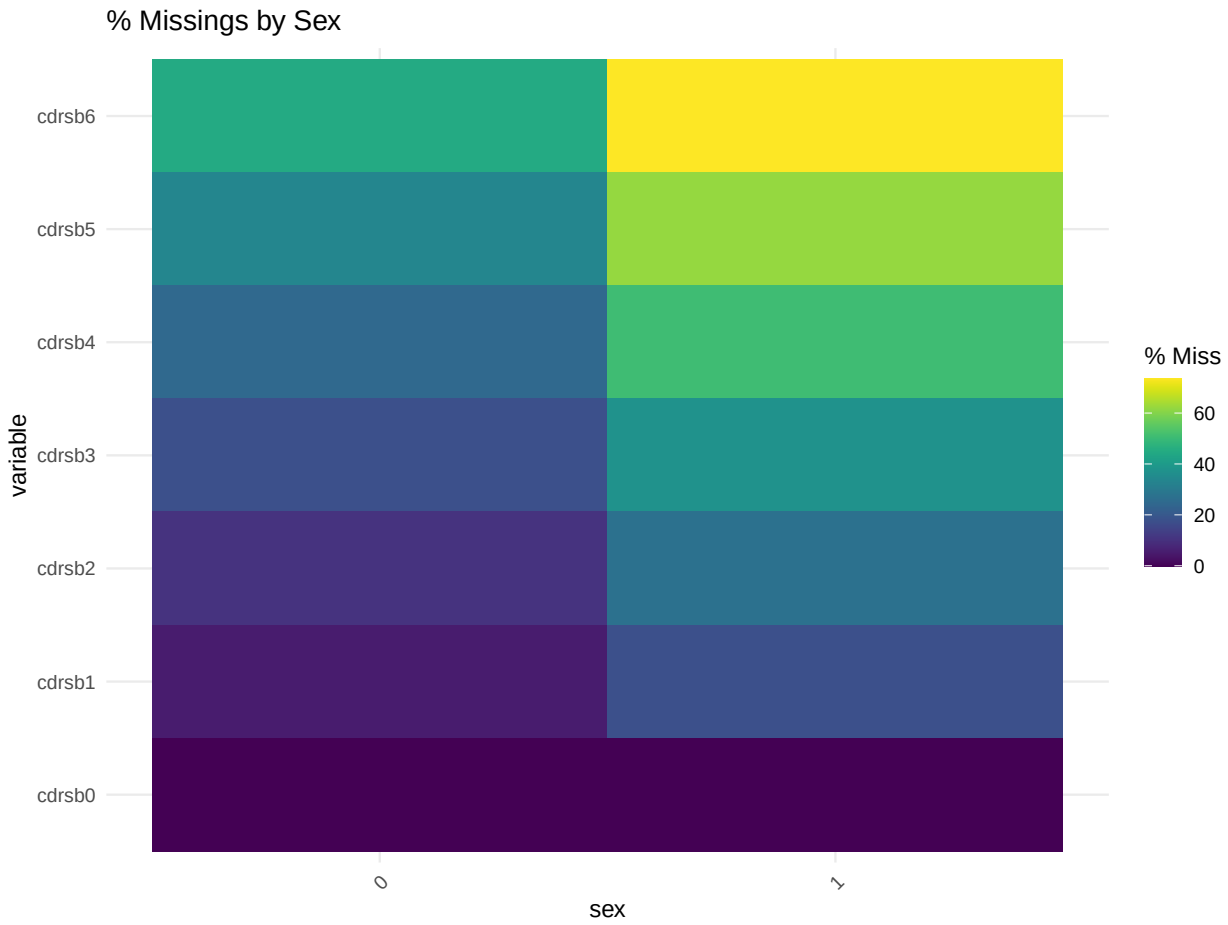
We conclude that missingness follows a **monotone** pattern.

4. Missingness per Covariates

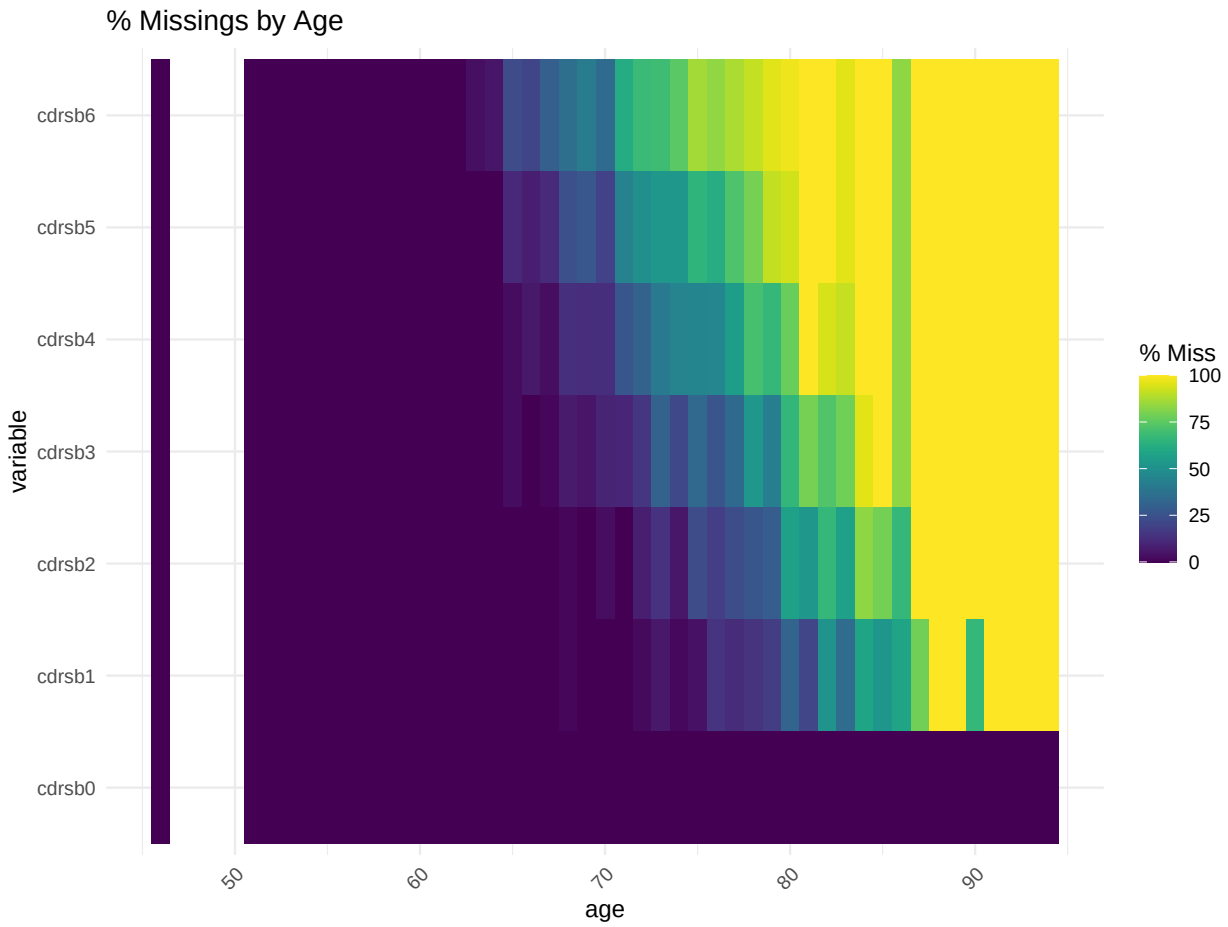
```
# Trial
alz %>% select(trial, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = trial) + labs(title = "% Missings by Trial")
```



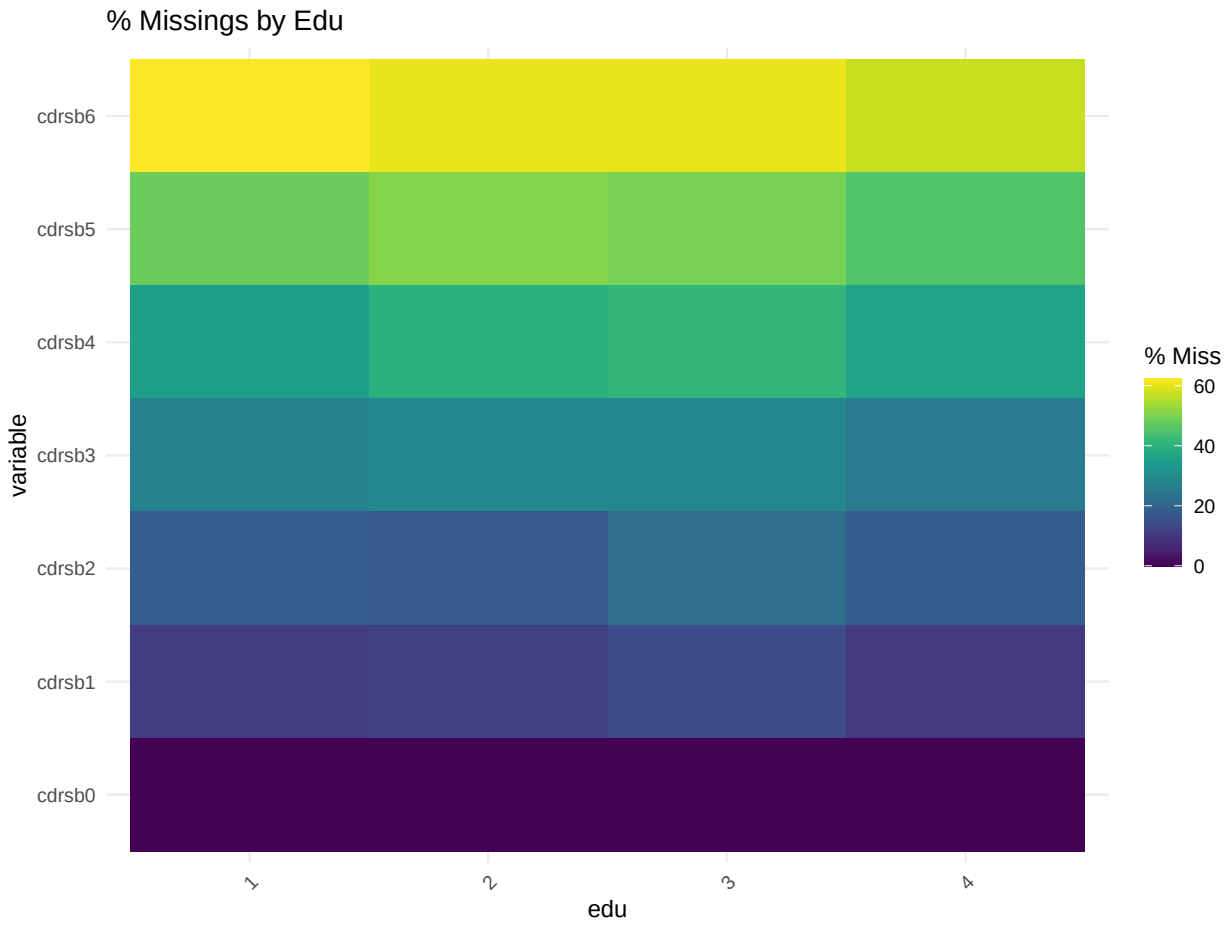

```
# Sex (0 - men, 1 - women)
alz %>% select(sex, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = sex) + labs(title = "% Missings by Sex")
```



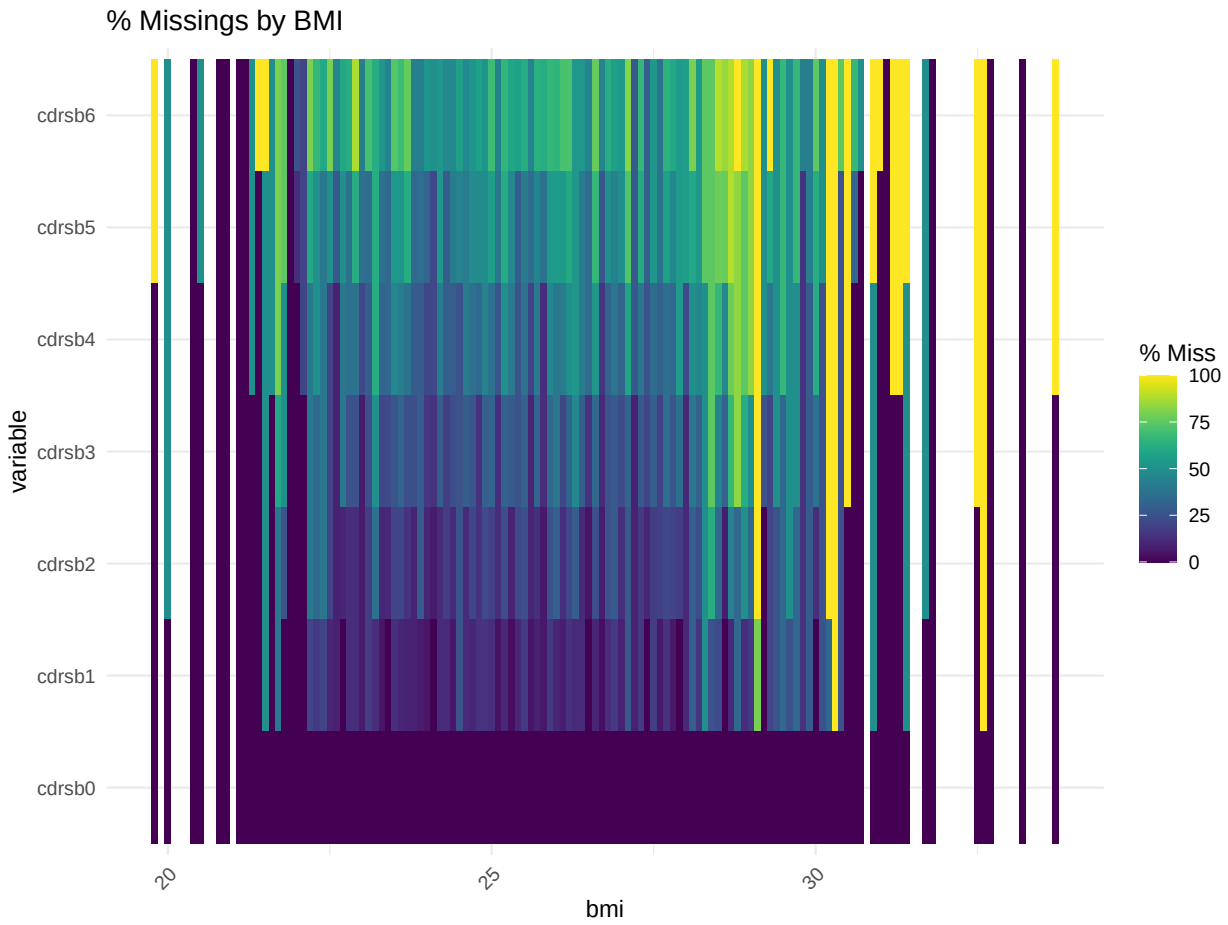

```
# Age
alz %>% select(age, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = age) + labs(title = "% Missings by Age")
```



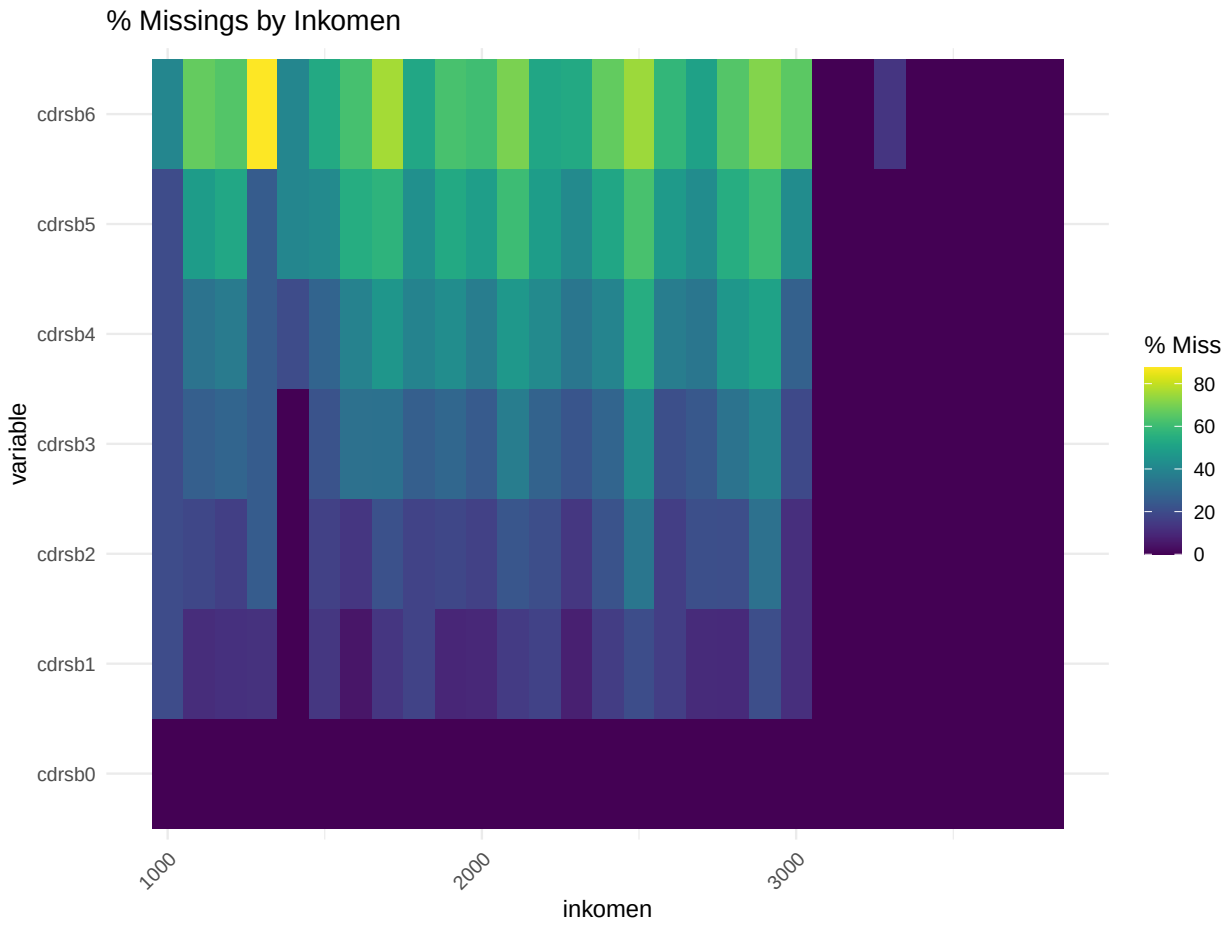

```
# Edu
alz %>% select(edu, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = edu) + labs(title = "% Missings by Edu")
```



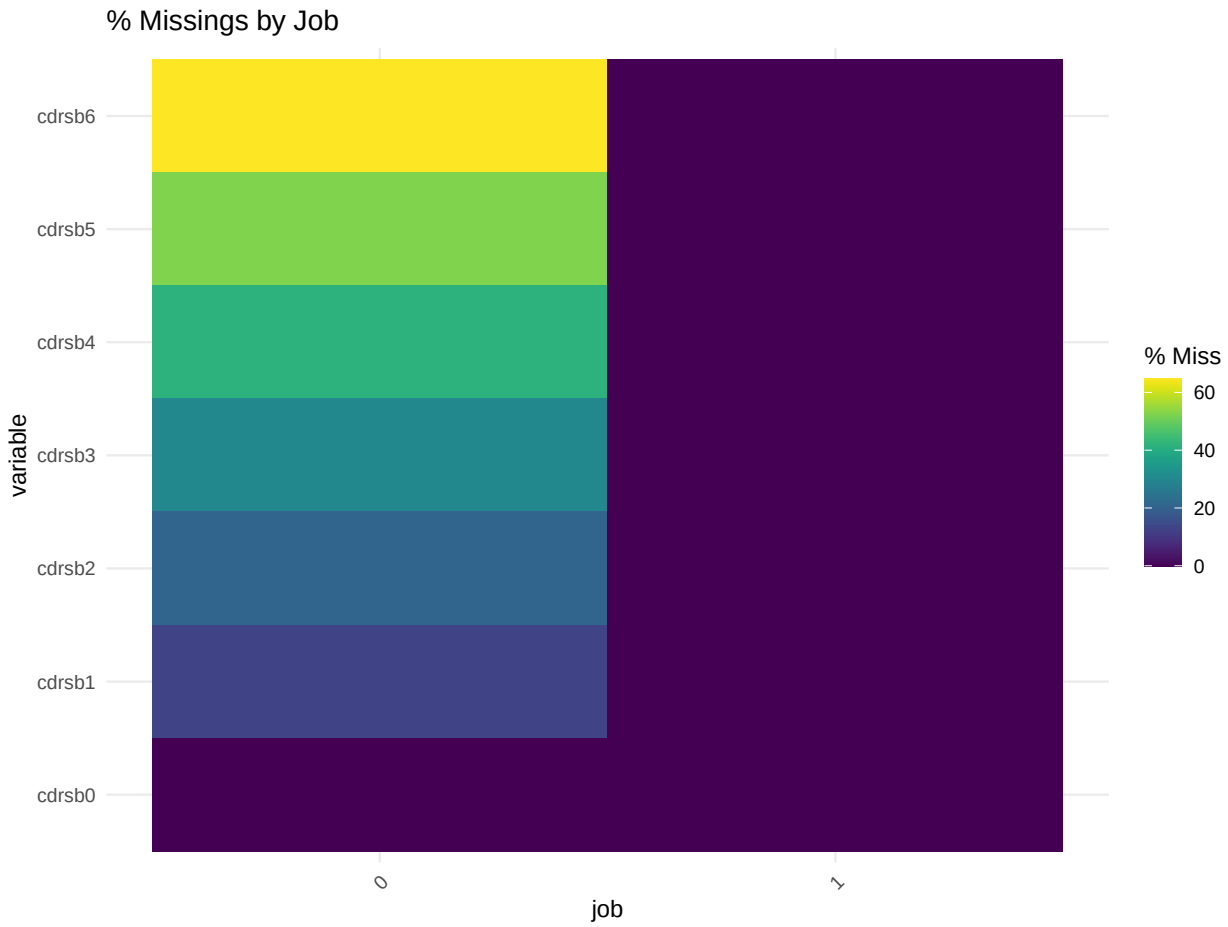

```
# BMI
alz %>% select(bmi, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = bmi) + labs(title = "% Missings by BMI")
```



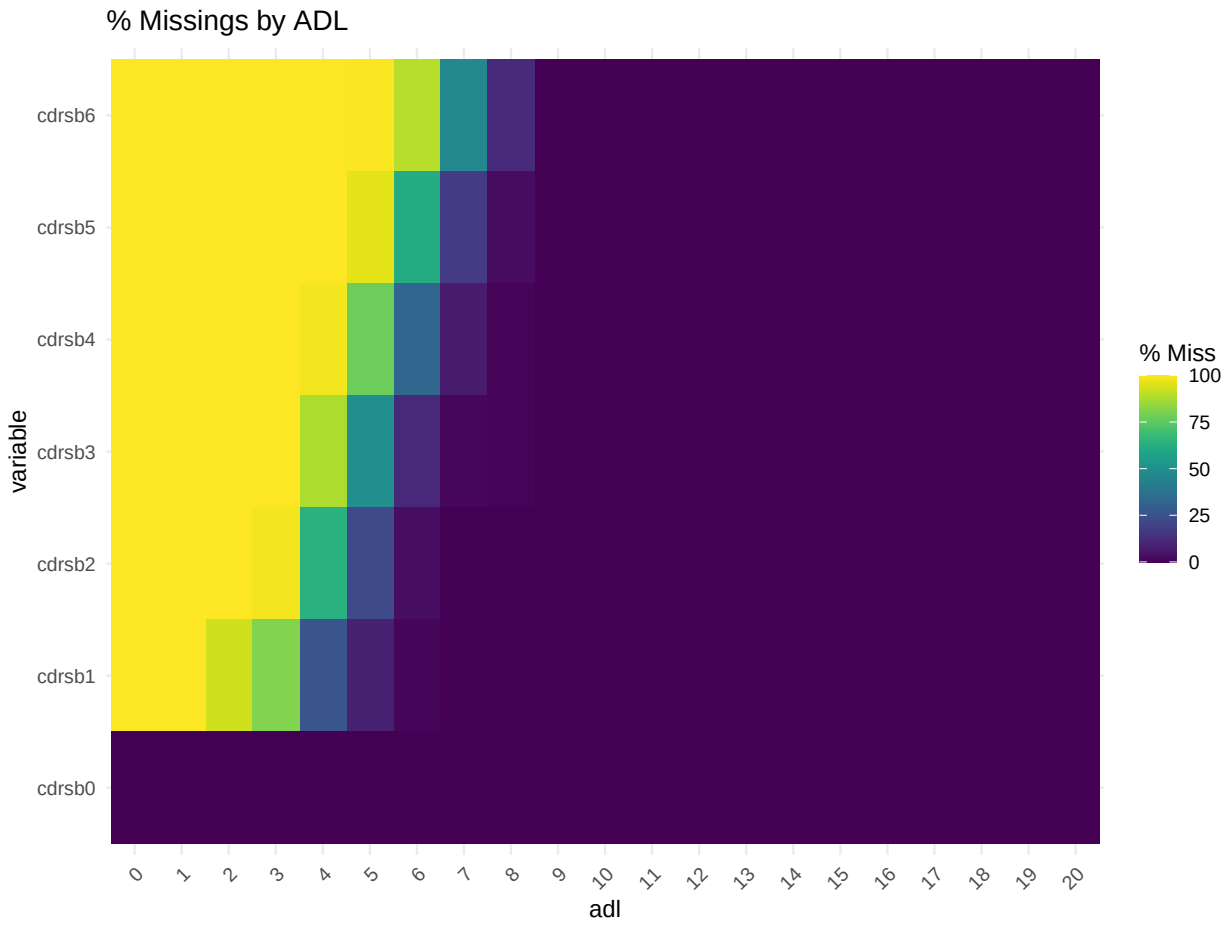

```
# Inkomen
alz %>% select(inkomen, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = inkomen) + labs(title = "% Missings by Inkomen")
```



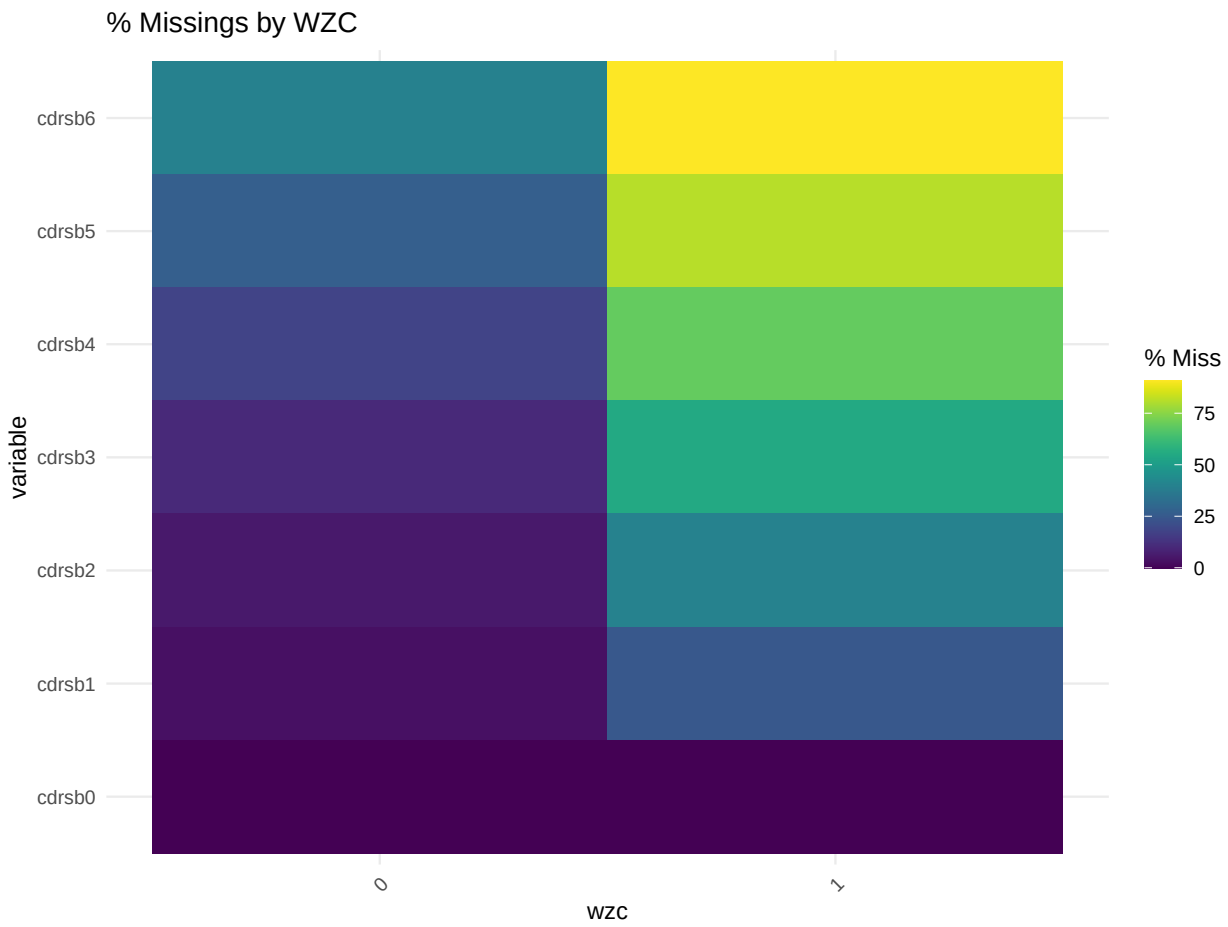

```
# Job
alz %>% select(job, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = job) + labs(title = "% Missings by Job")
```




```
# ADL
alz %>% select(adl, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = adl) + labs(title = "% Missings by ADL")
```



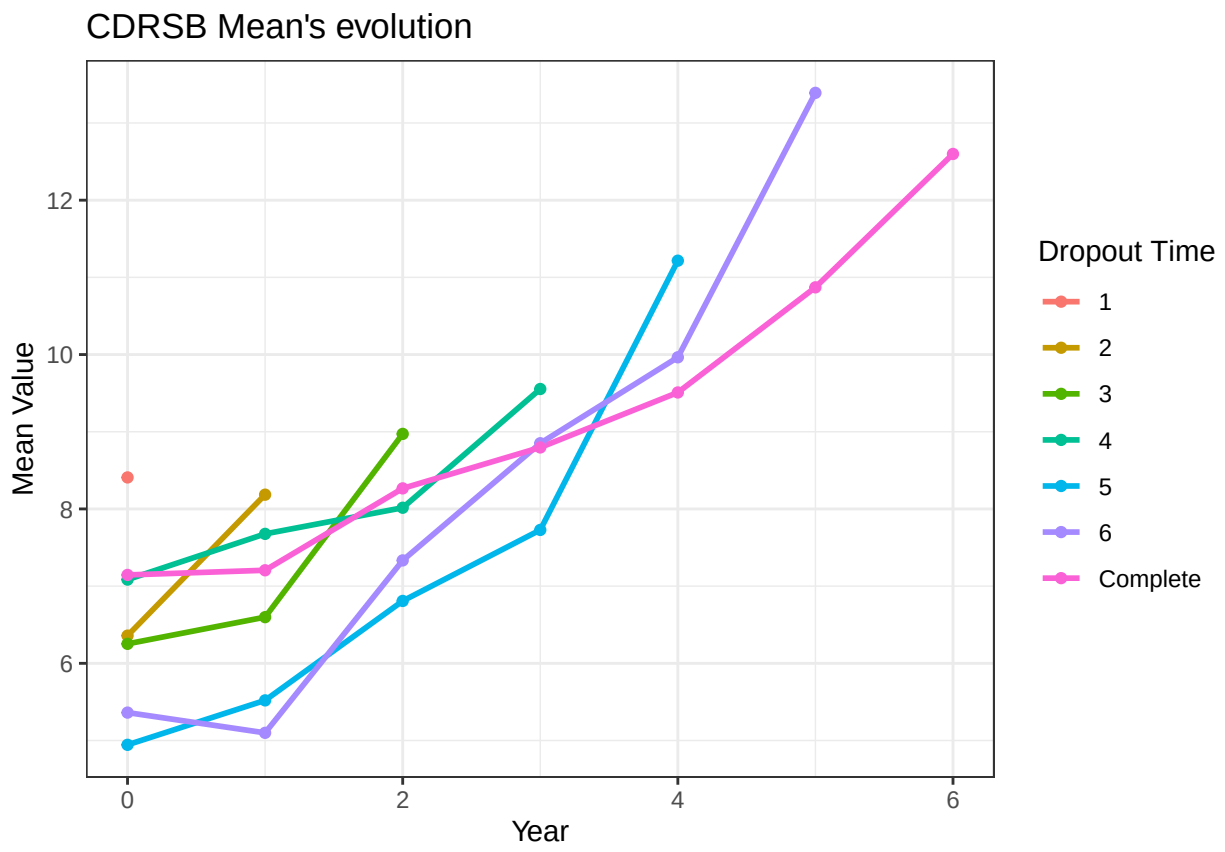

```
# WZC (0 - home, 1 - residence)
alz %>% select(wzc, matches("cdrsb\\d+$")) %>%
  gg_miss_fct(fct = wzc) + labs(title = "% Missings by WZC")
```



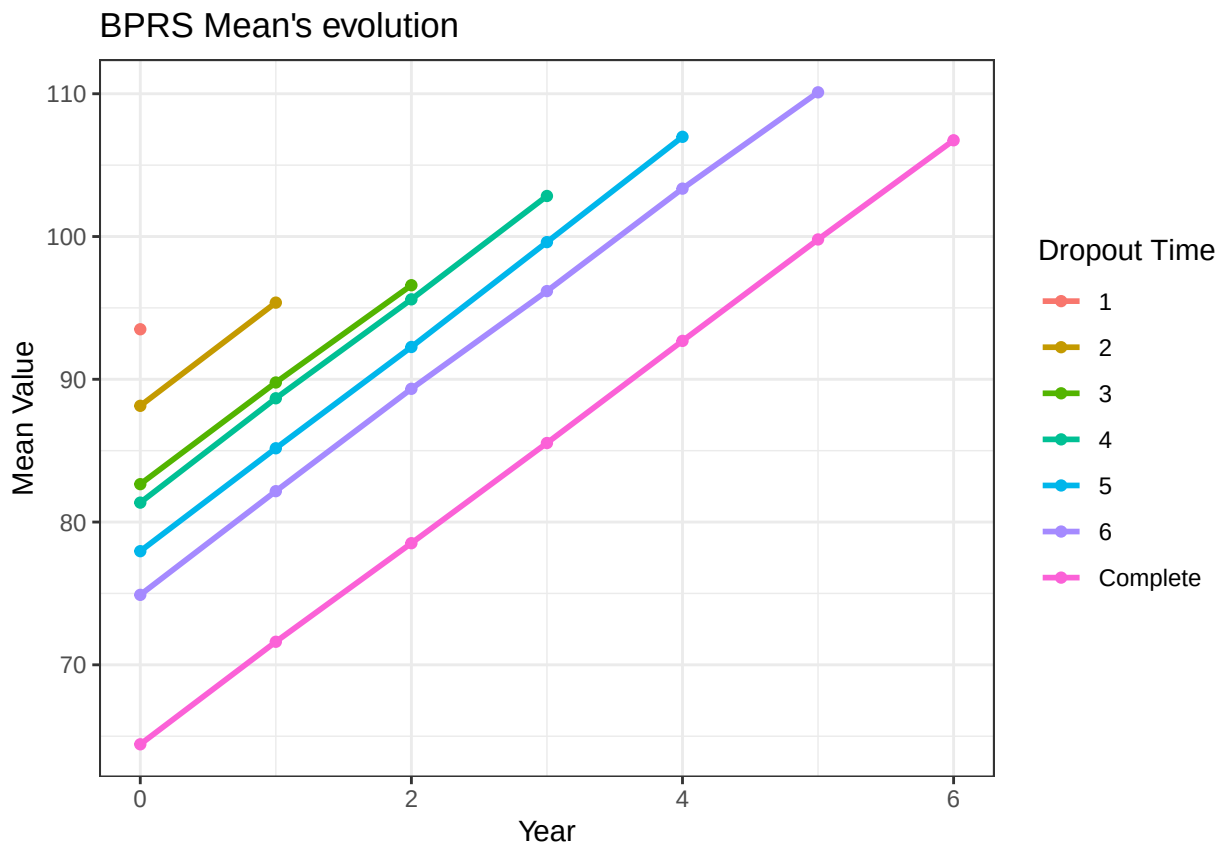
5. Mean Profiles by Dropout Time

```
dropout_info <- alz_long %>%
  group_by(id) %>%
  summarise(
    first_missing = ifelse(any(is.na(cdrsb)),
                           as.character(min(time[is.na(cdrsb)])),
                           "Complete")
  )

alz_patterns <- left_join(alz_long, dropout_info, by = "id")

# Function to reuse code
plot_mean_profile <- function(data, var, title) {
  data %>%
    group_by(time, first_missing) %>%
    summarise(mean_val = mean({{var}}, na.rm = TRUE), .groups = 'drop') %>%
    ggplot(aes(x = time, y = mean_val, color = as.factor(first_missing))) +
    geom_line(size = 1) +
    geom_point() +
    labs(title = title, x = "Year", y = "Mean Value", color = "Dropout Time") +
    theme_bw()
```

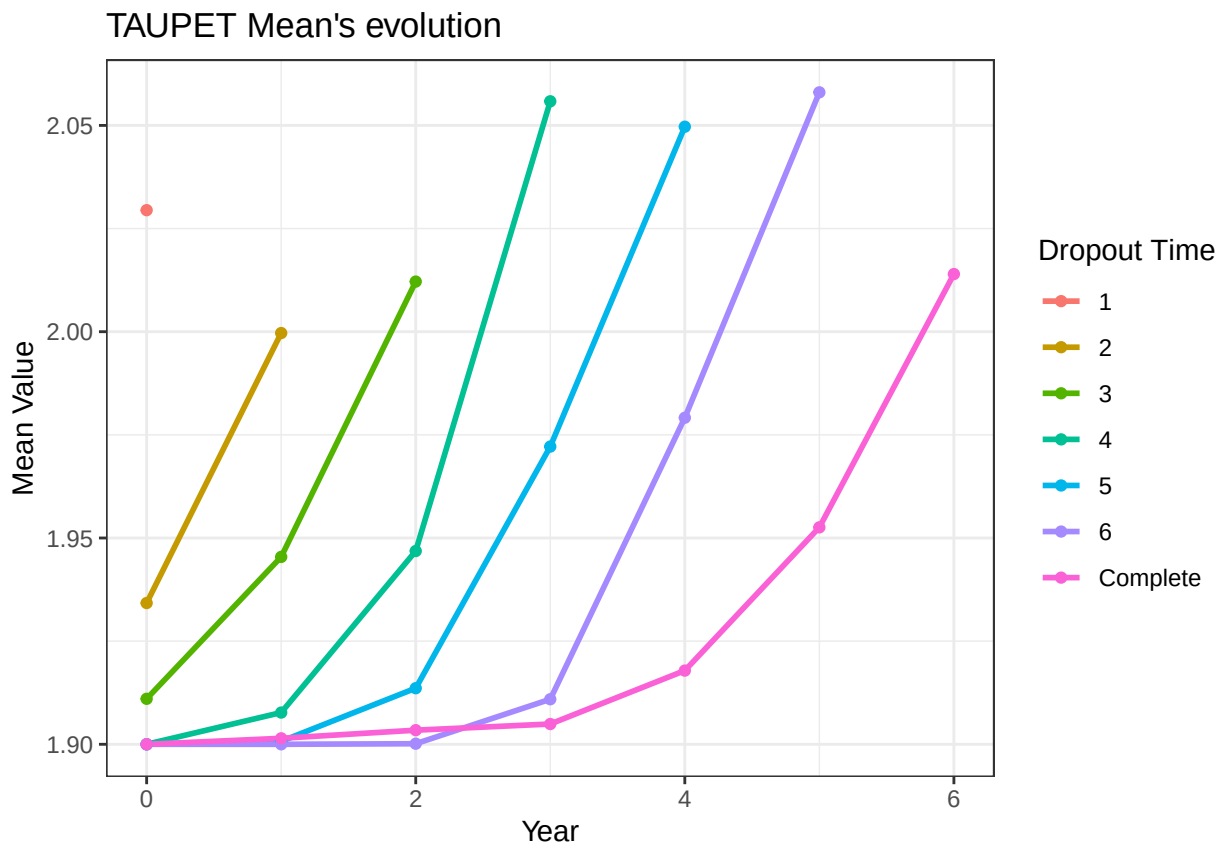
```
}

plot_mean_profile(alz_patterns, cdrsb, "CDRSB Mean's evolution")
```



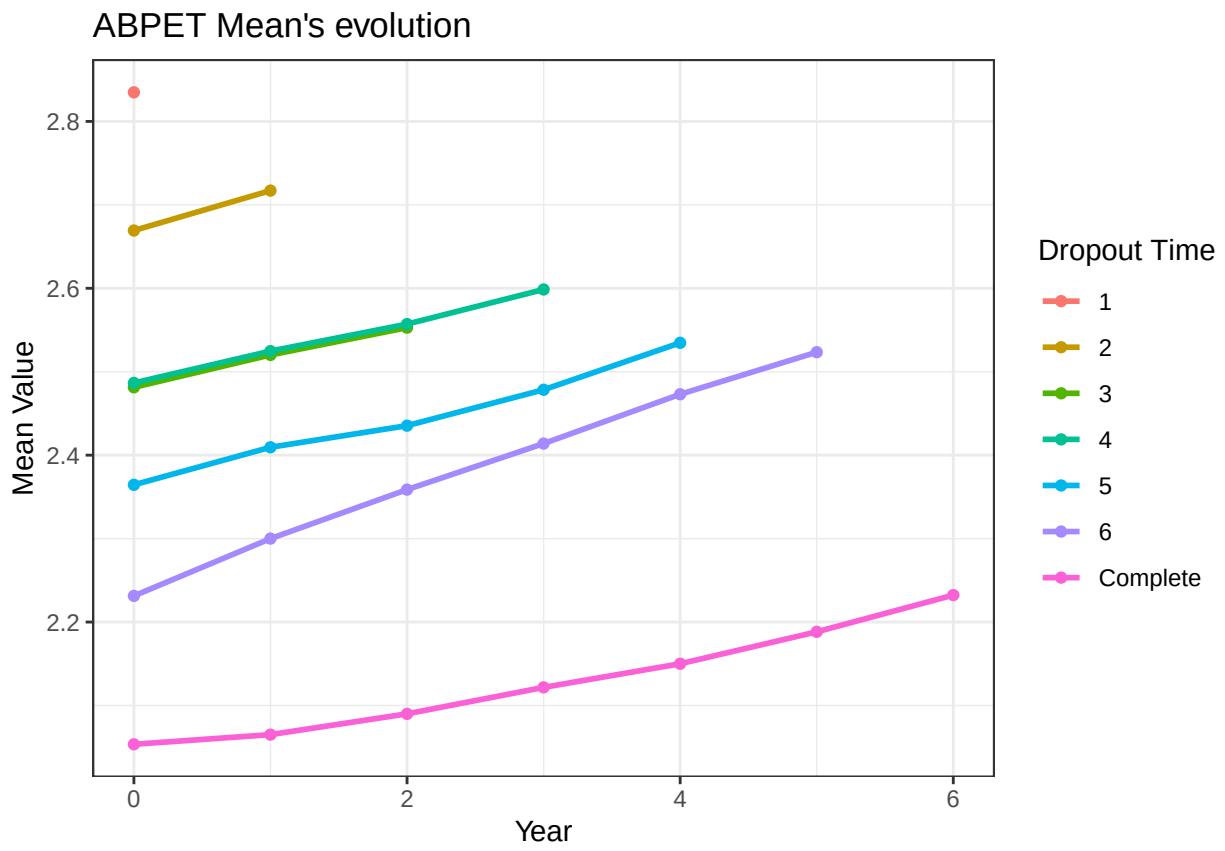

```
plot_mean_profile(alz_patterns, bprs, "BPRS Mean's evolution")
```




```
plot_mean_profile(alz_patterns, taupet, "TAUPET Mean's evolution")
```




```
plot_mean_profile(alz_patterns, abpet, "ABPET Mean's evolution")
```

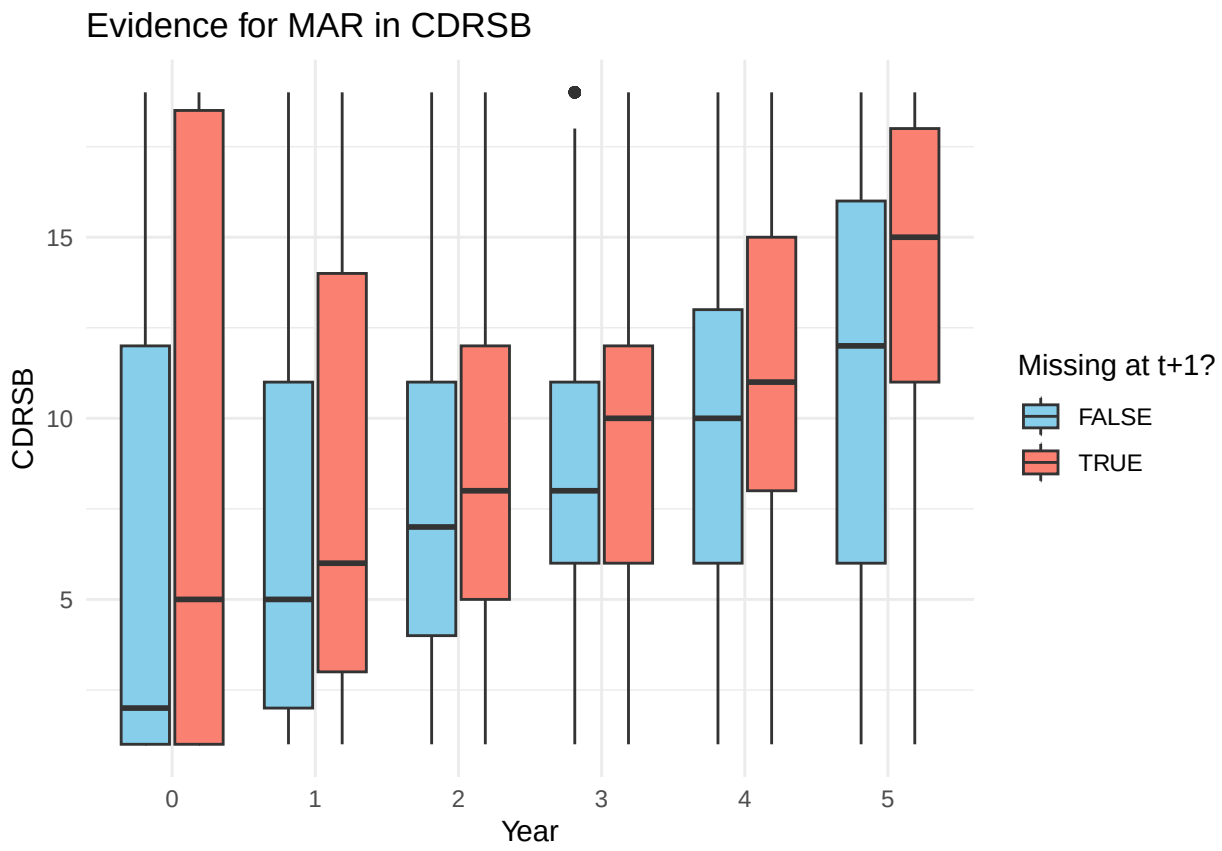


This may suggest us that we do **not** have a **MCAR** case, because patients who drop out have higher means (worse condition) than the completers.

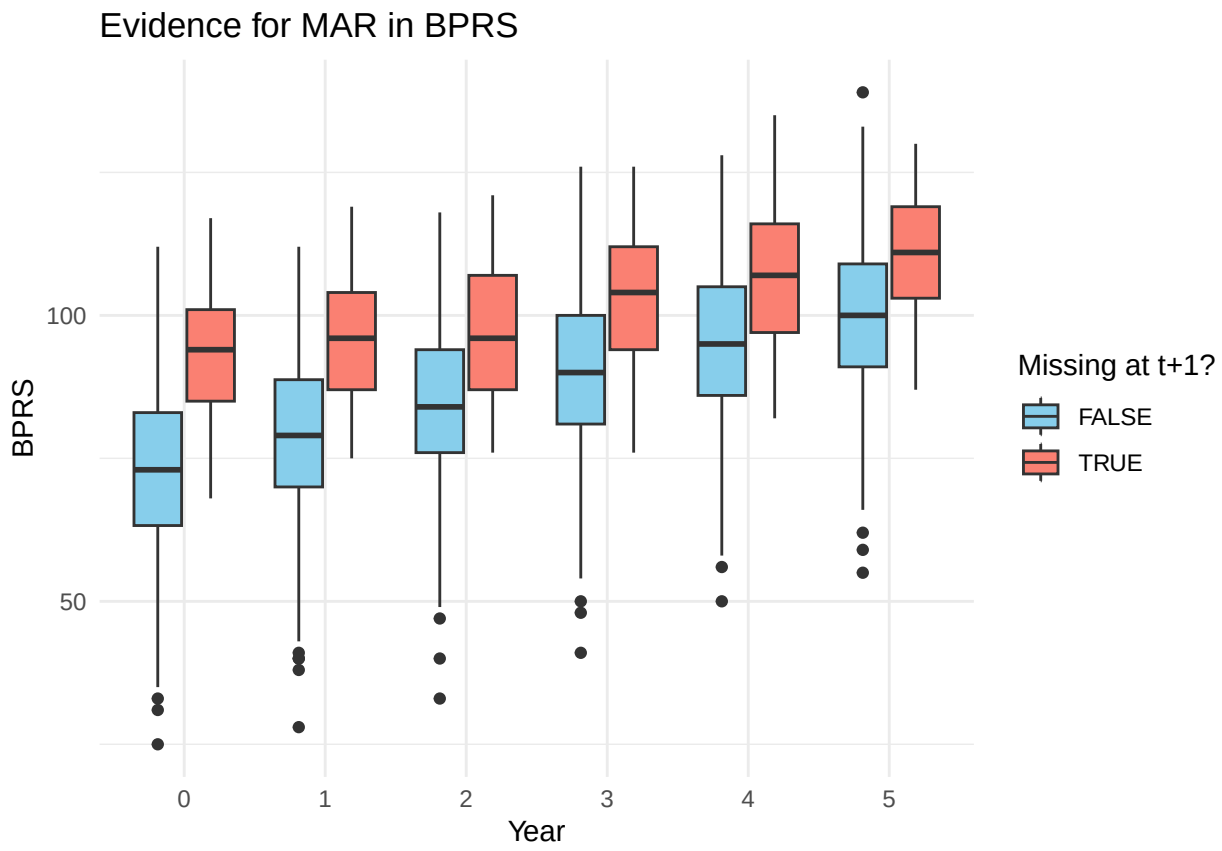
6. Dropout Mechanism Analysis

Evidence for MAR (Boxplots)

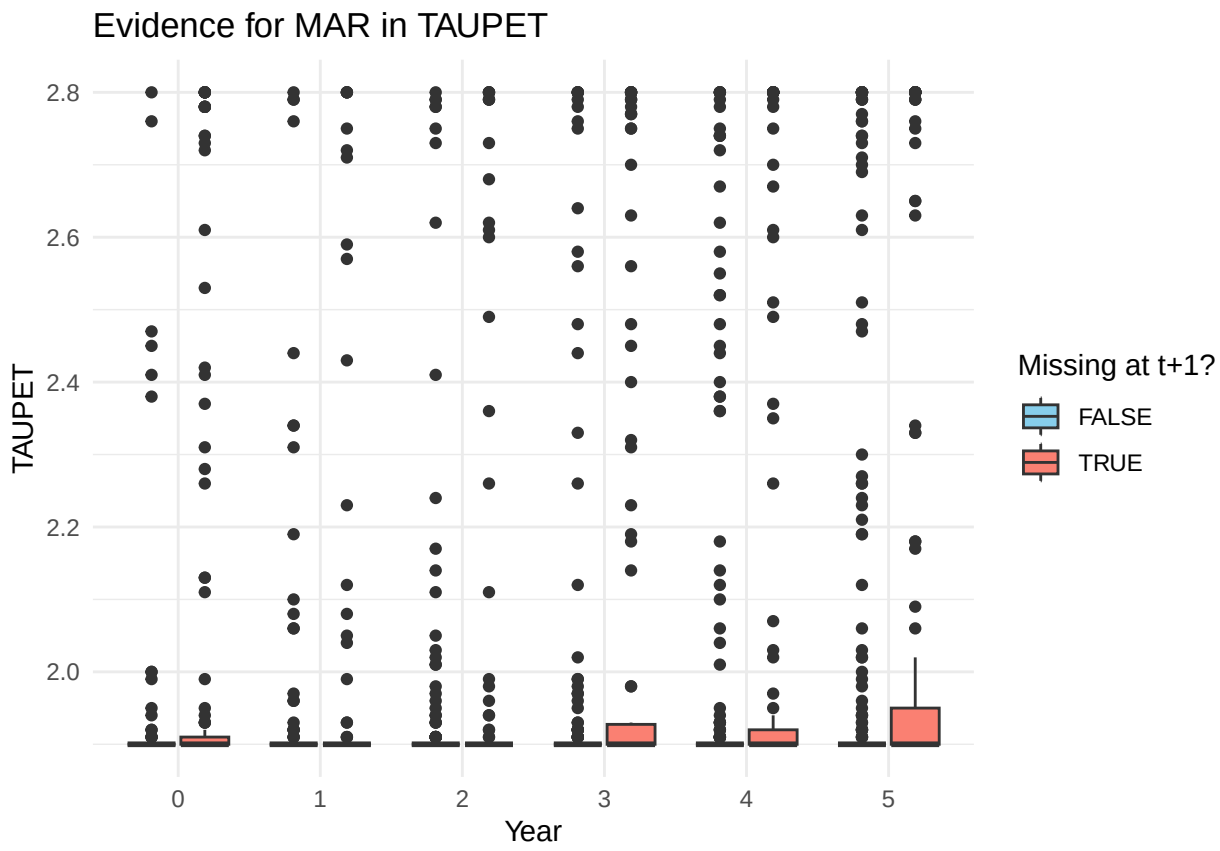
```
alz_long_lag <- alz_long %>%
  group_by(id) %>%
  mutate(
    is_missing_next = lead(is.na(cdrsb))
  ) %>%
  filter(time < max(time))

# Boxplot CDRSB
ggplot(alz_long_lag, aes(x = as.factor(time), y = cdrsb, fill = is_missing_next)) +
  geom_boxplot() +
  labs(title = "Evidence for MAR in CDRSB", x = "Year", y = "CDRSB", fill = "Missing at t+1?") +
  scale_fill_manual(values = c("skyblue", "salmon"), na.value="grey") +
  theme_minimal()
```



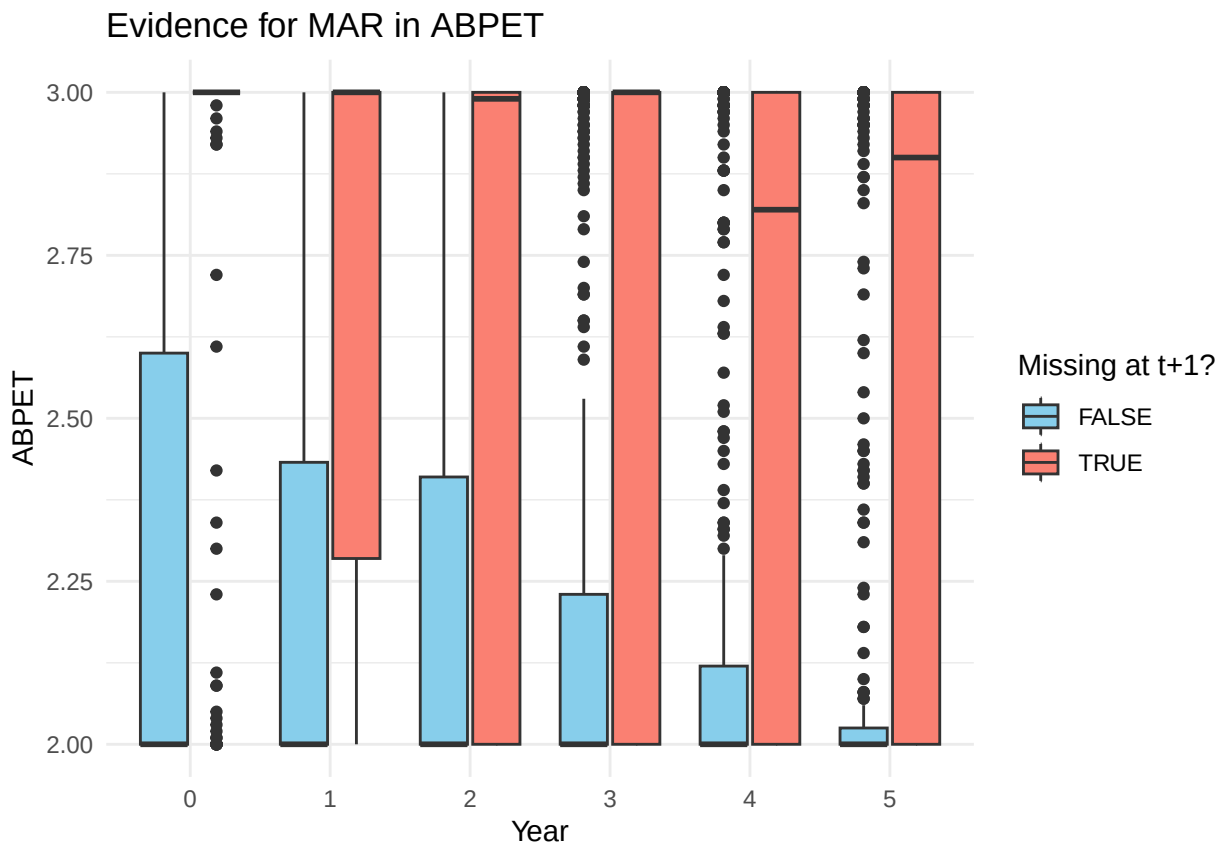

```
# Boxplot BPRS
ggplot(alz_long_lag, aes(x = as.factor(time), y = bprs, fill = is_missing_next)) +
  geom_boxplot() +
  labs(title = "Evidence for MAR in BPRS", x = "Year", y = "BPRS", fill = "Missing at t+1?") +
  scale_fill_manual(values = c("skyblue", "salmon"), na.value="grey") +
  theme_minimal()
```




```
# Boxplot TAUPET
ggplot(alz_long_lag, aes(x = as.factor(time), y = taupet, fill = is_missing_next)) +
  geom_boxplot() +
  labs(title = "Evidence for MAR in TAUPET", x = "Year", y = "TAUPET", fill = "Missing at t+1?") +
  scale_fill_manual(values = c("skyblue", "salmon"), na.value="grey") +
  theme_minimal()
```




```
# Boxplot ABPET
ggplot(alz_long_lag, aes(x = as.factor(time), y = abpet, fill = is_missing_next)) +
  geom_boxplot() +
  labs(title = "Evidence for MAR in ABPET", x = "Year", y = "ABPET", fill = "Missing at t+1?") +
  scale_fill_manual(values = c("skyblue", "salmon"), na.value="grey") +
  theme_minimal()
```



This suggests **MAR**, because the red boxes are higher than the blue ones. The patients that are worse leave the study.

Logistic Dropout Model

```
alz_dropout_model <- alz_long %>%
  group_by(id) %>%
  mutate(
    is_missing = is.na(cdrsb),
    prev_cdrsb = lag(cdrsb),
    prev_bprs = lag(bprs) ,
    prev_taupet = lag(taupet),
    prev_abpet = lag(abpet)
  ) %>%
  filter(time > 0) %>%
  ungroup()

model_dropout <- glm(is_missing ~ prev_cdrsb + prev_bprs + prev_taupet + prev_abpet +
                                  trial + sex + age + edu + bmi + inkomen + job + adl_num + wzc + time,
                     data = alz_dropout_model,
                     family = binomial)

summary(model_dropout)
```

```
##
```

```
## Call:
## glm(formula = is_missing ~ prev_cdrsb + prev_bprs + prev_taupet +
##     prev_abpet + trial + sex + age + edu + bmi + inkomen + job +
##     adl_num + wzc + time, family = binomial, data = alz_dropout_model)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.638e+00  6.611e+00  -0.550   0.5821
## prev_cdrsb   1.209e-01  1.103e-02  10.964   <2e-16 ***
## prev_bprs    1.331e-03  1.785e-02   0.075   0.9406
## prev_taupet  1.084e-01  3.350e-01   0.324   0.7462
## prev_abpet  -8.982e-02  2.718e-01  -0.331   0.7410
## trial2       6.834e-01  5.665e-01   1.206   0.2277
## trial3       2.631e-01  5.416e-01   0.486   0.6272
## trial4      -6.377e-01  6.424e-01  -0.993   0.3209
## trial5       1.483e-01  5.299e-01   0.280   0.7796
## trial6       4.473e-01  6.650e-01   0.673   0.5012
## trial7      -8.951e-02  5.142e-01  -0.174   0.8618
## trial8       5.666e-01  5.683e-01   0.997   0.3187
## trial9      -5.234e-02  4.927e-01  -0.106   0.9154
## trial10      6.270e-01  6.833e-01   0.918   0.3589
## trial11     -4.496e-02  5.425e-01  -0.083   0.9340
## trial12      2.856e-01  5.516e-01   0.518   0.6047
## trial13      5.913e-01  5.250e-01   1.126   0.2601
## trial14      2.555e-01  6.253e-01   0.409   0.6829
## trial15      5.073e-01  5.680e-01   0.893   0.3718
## trial16      4.439e-01  6.664e-01   0.666   0.5054
## trial17      4.639e-02  5.430e-01   0.085   0.9319
## trial18     -1.024e-01  6.557e-01  -0.156   0.8759
## trial19      1.876e-01  5.009e-01   0.374   0.7081
## trial20     -1.443e-02  6.112e-01  -0.024   0.9812
## trial21     -1.390e-01  5.970e-01  -0.233   0.8158
## trial22      3.269e-01  4.941e-01   0.662   0.5082
## trial23     -2.847e-01  6.306e-01  -0.451   0.6517
## trial24     -1.519e-01  5.676e-01  -0.268   0.7890
## trial25      9.563e-02  6.616e-01   0.145   0.8851
## sex1         4.562e-01  2.672e-01   1.708   0.0877 .
## age          8.632e-02  6.948e-02   1.242   0.2141
## edu2         3.942e-01  2.864e-01   1.377   0.1687
## edu3         1.029e+00  4.250e-01   2.421   0.0155 *
## edu4         1.079e+00  6.080e-01   1.774   0.0761 .
## bmi          7.651e-02  4.177e-02   1.832   0.0670 .
## inkomen     -6.046e-04  4.824e-04  -1.253   0.2101
## job1        -8.967e+00  2.777e+02  -0.032   0.9742
## adl_num     -1.971e+00  1.911e-01 -10.315   <2e-16 ***
## wzc1         1.984e+00  1.441e-01  13.766   <2e-16 ***
## time         1.441e+00  1.421e-01  10.140   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4411.1  on 5708  degrees of freedom
## Residual deviance: 1943.5  on 5669  degrees of freedom
```

```
##   (1809 observations deleted due to missingness)
## AIC: 2023.5
##
## Number of Fisher Scoring iterations: 18
```

Since we have that the p-value of `prev_cdrsb` < 2e-16*, we have evidence of **MAR****. Since the estimate for `prev_cdrsb` = 0.1209 > 0, patients with higher cdrsb tend to dropout more.

MCAR Tests (Little's Test)

```
mcar_cdrsb <- mcar_test(alz %>% select(matches("cdrsb\\d+$")))
mcar_bprs <- mcar_test(alz %>% select(matches("bprs\\d+$")))
mcar_taupet <- mcar_test(alz %>% select(matches("taupet\\d+$")))
mcar_abpet <- mcar_test(alz %>% select(matches("abpet\\d+$")))

print(mcar_cdrsb)
```

```
## # A tibble: 1 x 4
##   statistic    df       p.value missing.patterns
##       <dbl> <dbl>         <dbl>            <int>
## 1      82.5    21 0.00000000310                7
```

```
print(mcar_bprs)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      616.    21       0                7
```

```
print(mcar_taupet)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      352.    21       0                7
```

```
print(mcar_abpet)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      521.    21       0                7
```

Since the p-values of these tests are ~ 0, we conclude that we do **not** have a **MCAR** case.

7. Baseline Comparison: Completers vs Dropouts

```r
completers_id <- alz_long %>%
  filter(time == 6 & !is.na(cdrsb)) %>%
  pull(patid)

alz_comparison <- alz %>%
  mutate(
    Status = ifelse(patid %in% completers_id, "Completer", "Dropout")
  )

comp_table <- alz_comparison %>%
  group_by(Status) %>%
  summarise(
    N = n(),
    CDRSB_Base = mean(cdrsb0, na.rm=TRUE),
    BPRS_Base = mean(bprs0, na.rm=TRUE),
    Age_Mean = mean(age, na.rm=TRUE),
    ADL_Mean = mean(adl_num, na.rm = TRUE)
  )

knitr::kable(comp_table, digits = 2)
```

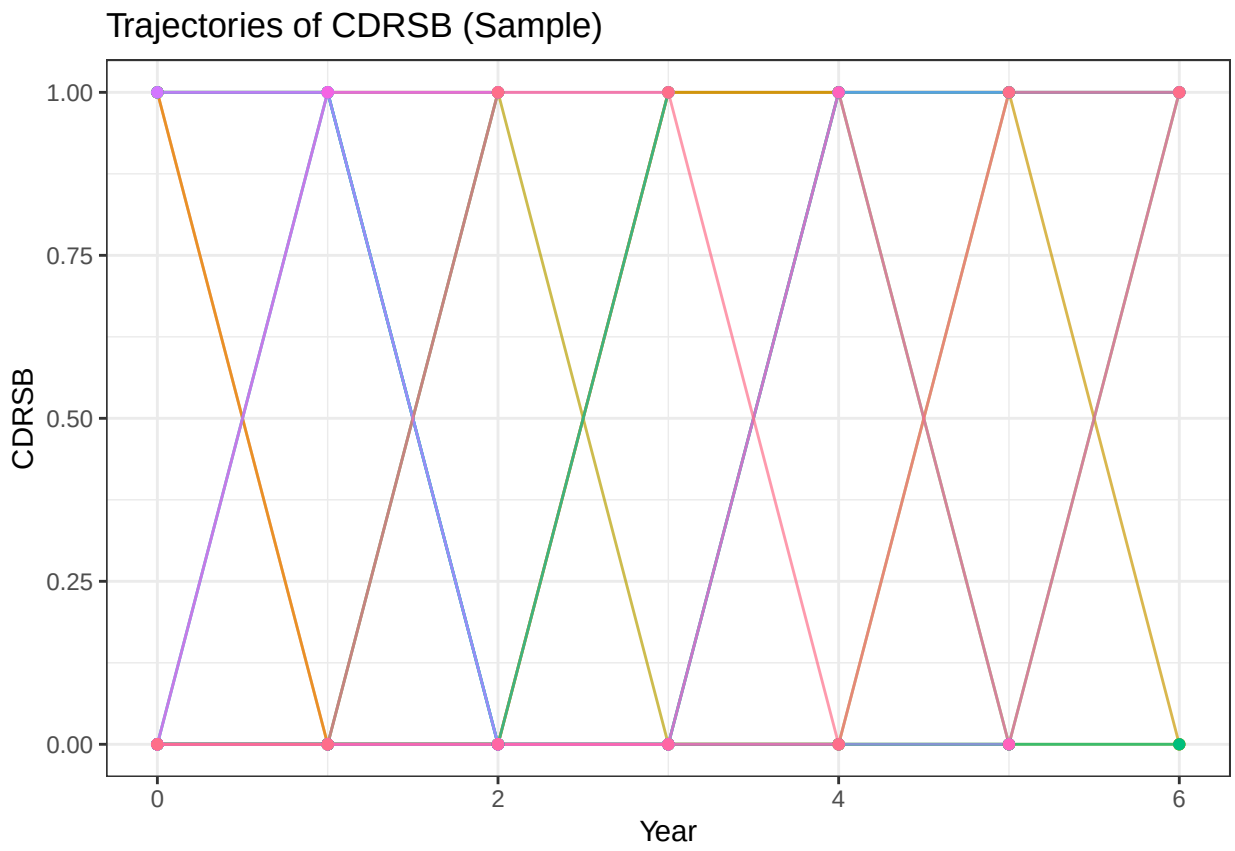
Status	N	CDRSB_Base	BPRS_Base	Age_Mean	ADL_Mean
Completer	511	7.14	64.43	66.75	10.51
Dropout	742	6.45	82.99	76.38	6.04

The descriptive comparison reveals that the **dropout group** had a higher mean Age and higher baseline BPRS scores, while conversely presenting a lower mean baseline CDRSB and lower ADL scores.

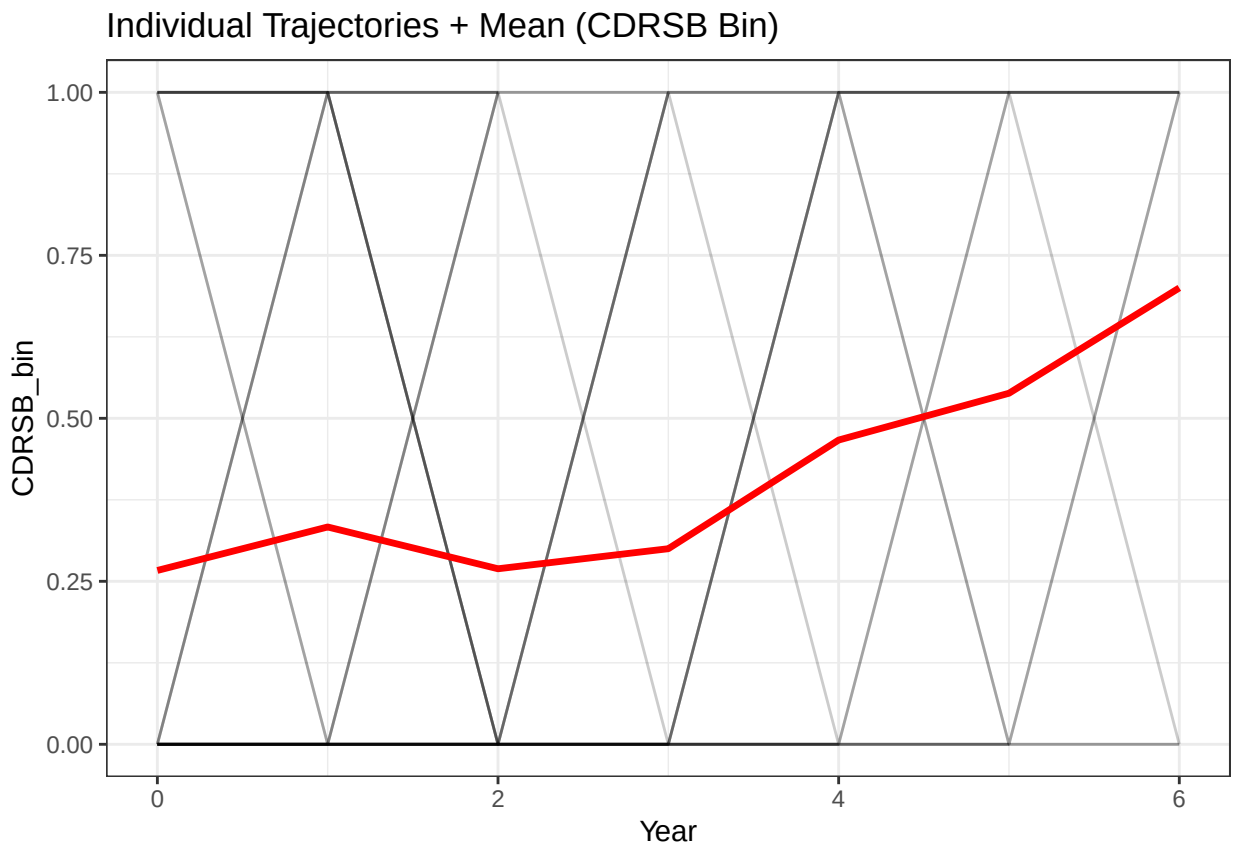
8. Additional Visualizations

Spaghetti Plots

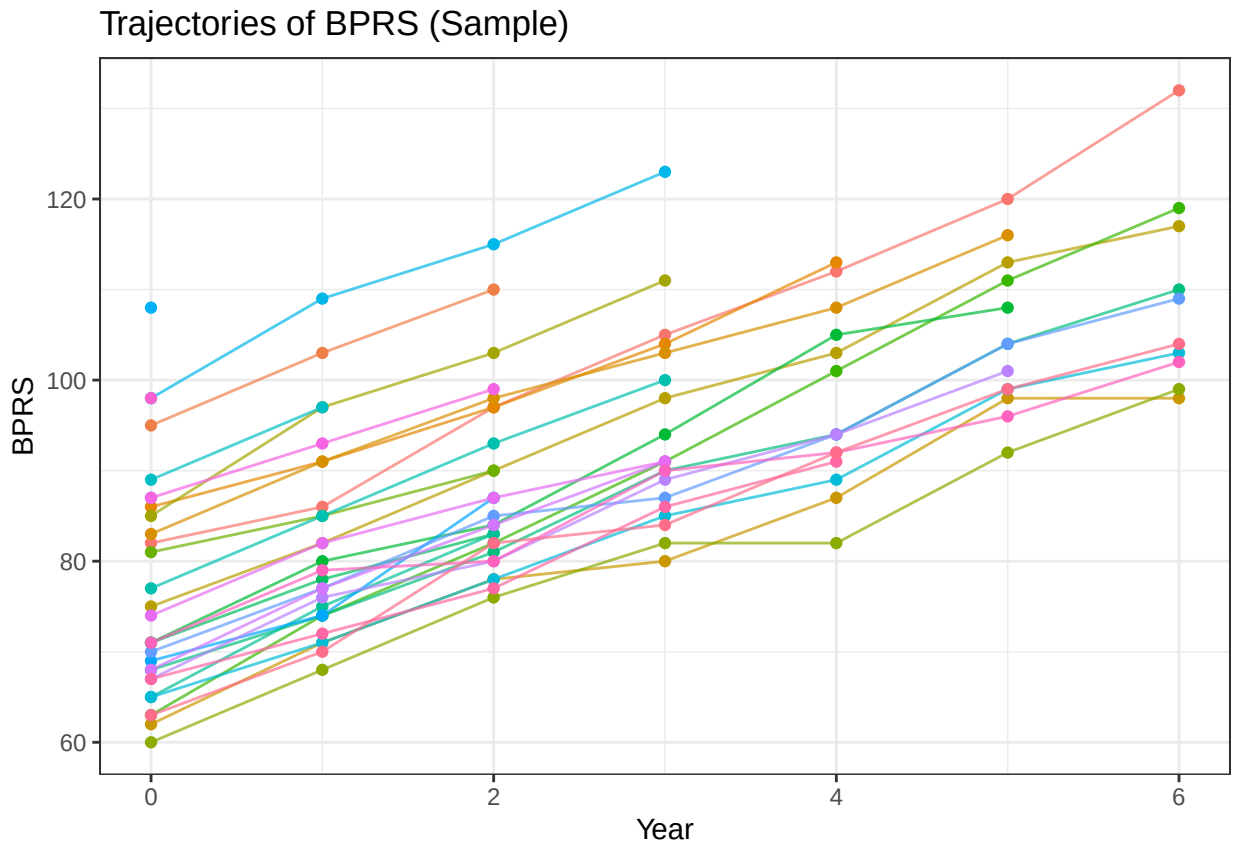
```r
set.seed(123)
random_ids <- sample(unique(alz_long$patid), 30)

sample_data <- alz_long %>%
  filter(patid %in% random_ids)

# CDRSB Individual
ggplot(sample_data, aes(x = time, y = cdrsb_bin, group = patid, color = as.factor(patid))) +
  geom_line(alpha = 0.7) +
  geom_point() +
  theme_bw() +
  theme(legend.position = "none") +
  labs(title = "Trajectories of CDRSB (Sample)", y = "CDRSB", x = "Year")
```




```
# CDRSB Mean Trend
ggplot(sample_data, aes(x = time, y = cdrsb_bin, group = patid)) +
  geom_line(alpha = 0.2) +
  stat_summary(aes(group = 1), fun = mean, geom = "line", color = "red", size = 1.2) +
  labs(title = "Individual Trajectories + Mean (CDRSB Bin)",
       y = "CDRSB_bin", x = "Year") +
  theme_bw()
```

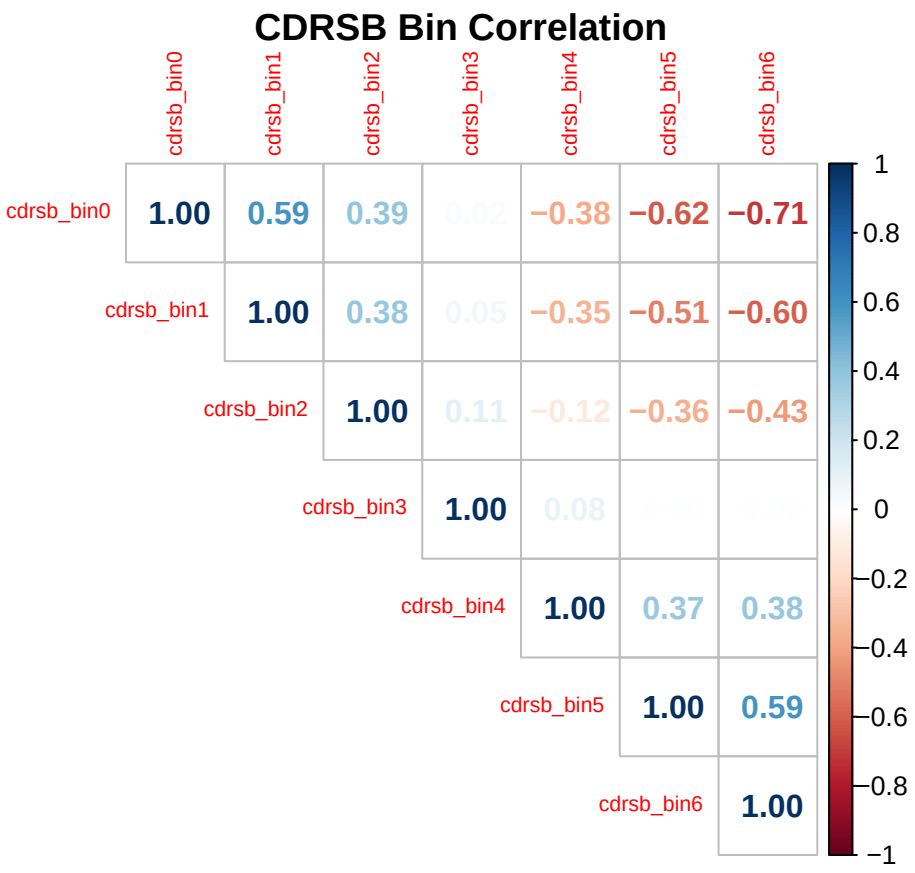



```
# BPRS Individual
ggplot(sample_data, aes(x = time, y = bprs, group = patid, color = as.factor(patid))) +
  geom_line(alpha = 0.7) +
  geom_point() +
  theme_bw() +
  theme(legend.position = "none") +
  labs(title = "Trajectories of BPRS (Sample)", y = "BPRS", x = "Year")
```

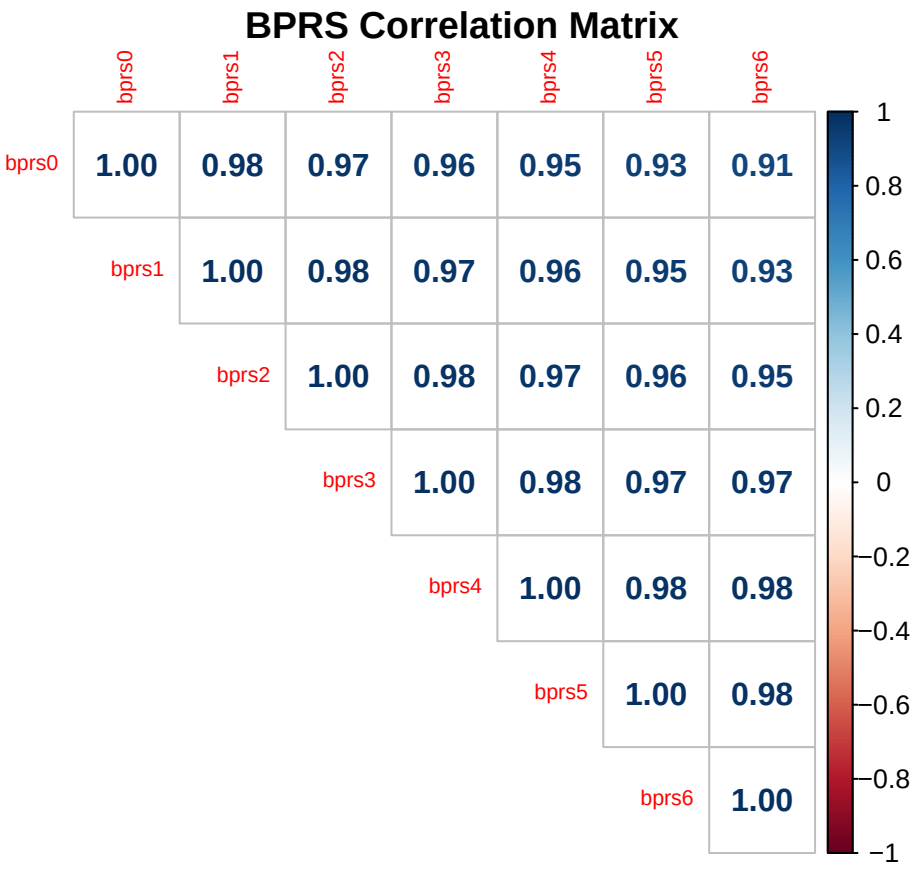


Correlation Matrices

```
# CDRSB
cor_data_cdrsb <- alz %>%
  select(matches("cdrsb_bin\\d+$")) %>%
  na.omit()

cor_matrix_cdrsb <- round(cor(cor_data_cdrsb), 2)
corrplot(cor_matrix_cdrsb,
         title = "CDRSB Bin Correlation",
         method = "number", type = "upper", tl.cex = 0.7, mar=c(0,0,1,0))
```

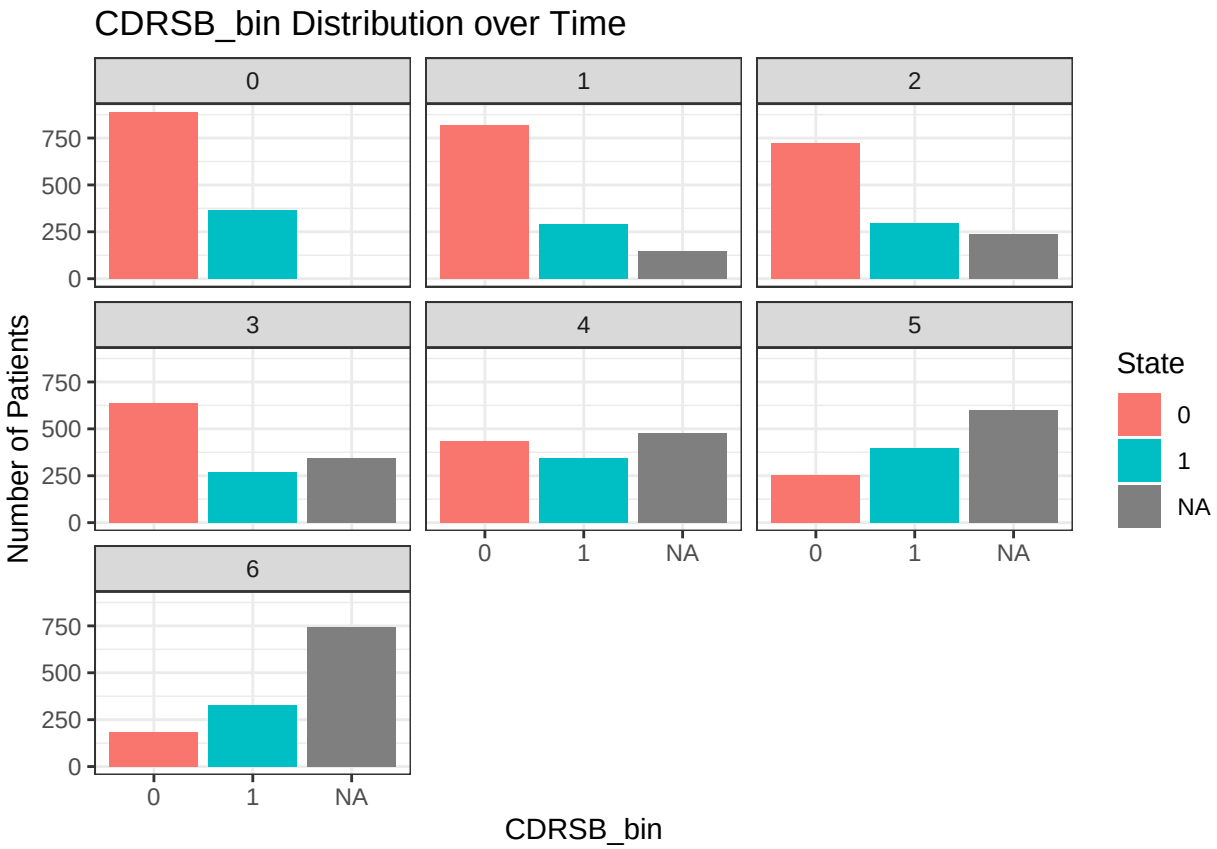



```r
# BPRS
cor_data_bprs <- alz %>%
  select(matches("bprs\\d+$")) %>%
  na.omit()

cor_matrix_bprs <- round(cor(cor_data_bprs), 2)
corrplot(cor_matrix_bprs,
         title = "BPRS Correlation Matrix",
         method = "number", type = "upper", tl.cex = 0.7, mar=c(0,0,1,0))
```

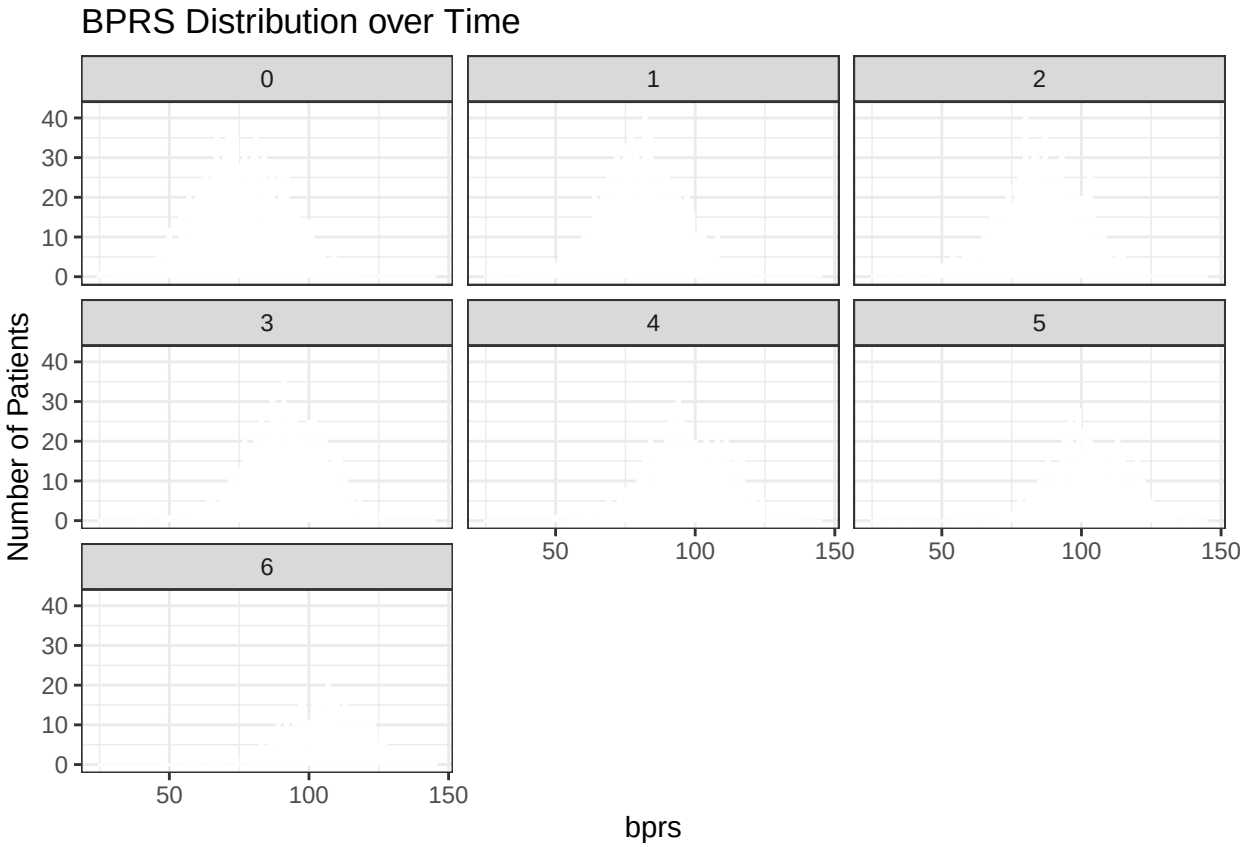


Distributions

```
# CDRSB Barplot
ggplot(alz_long, aes(x = as.factor(cdrsb_bin), fill = as.factor(cdrsb_bin))) +
  geom_bar() +
  facet_wrap(~time) +
  labs(title = "CDRSB_bin Distribution over Time",
       y = "Number of Patients",
       x = "CDRSB_bin",
       fill = "State") +
  theme_bw()
```




```
# BPRS Histogram
ggplot(alz_long, aes(x = bprs)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "white") +
  facet_wrap(~time) +
  labs(title = "BPRS Distribution over Time", y = "Number of Patients") +
  theme_bw()
```


BPRS Distribution over Time
0
1
2
3
4
5
6
Number of Patients
bprs
0
10
20
30
40
50
100
150