📊 Varsayımlar

📈 Kayıp Değerler

👩‍🏫 Dr. Kübra Atalay Kabasakal

Bahar 2023

İlk olarak veriyi okuma

```
library(haven)
screen <- read_sav("SCREEN.sav")
head(screen)
```

```
## # A tibble: 6 × 7
##   subno timedrs attdrug atthouse income mstatus  race
##   <dbl>   <dbl>   <dbl>    <dbl>  <dbl>   <dbl> <dbl>
## 1     1       1       8       27      5       2     1
## 2     2       3       7       20      6       2     1
## 3     3       0       8       23      3       2     1
## 4     4      13       9       28      8       2     1
## 5     5      15       7       24      1       2     1
## 6     6       3       8       25      4       2     1
```

Kayıp Veri

- Kayıp veri, veri analizindeki **en yaygın problemlerden** biridir.
- Kayıp verinin önemi **kayıp verinin miktarına, örüntüsüne ve neden eksik olduğuna** bağlıdır.
- Bir değişkene ait beklenmeyen miktarda kayıp veri varsa, ilk olarak bunun nedeni araştırılmalıdır. Daha sonra **kayıp verinin örüntüsüne bakılarak, rastlantısal mı yoksa sistematik bir örüntü mü gösterdiği belirlenmelidir.**
    - Örneğin, 30 yaşın üstündeki birçok kadın yaş ile ilgili soruyu cevaplamak istemezler.
- Genellikle kayıp verinin örüntüsü miktarından daha önemlidir. Rastlantısal dağılmayan kayıp veriler sonuçların genellenebilirliğini etkileyeceğinden miktarları az da olsa, **rastlantısal dağılan kayıp verilere oranla daha ciddi problemlere yol açarlar.**

Kayıp Veri Türleri

- Kayıp veri türleri arasındaki ayrım 1976 yılında Rubin tarafından yapılmıştır. Rubin (1976) kayıp veriyi aşağıdaki şekilde sınıflandırmıştır.
  - **Tamamen Rastlantısal Olarak Kayıp (TROK)** Missing Completely at R andom **MCAR**
  - **Rastlantısal Olarak Kayıp (ROK)** Missing at R andom **MAR**
  - **Rastlantısal Olmayarak Kayıp**/İhmal Edilemez Kayıp (İEK) Not Missing at R andom **NMAR**
- Kayıp veri en azından MAR türünde değilse, kayıp verinin **ihmal edilemeyeceği söylenir** ve bu türdeki kayıp veri **rastlantısal olmayarak kayıp veya ihmal edilemez kayıp olarak adlandırılır.**

- Rubin, D. B. (1976). Inference with missing data. Biometrika , 63, 581 592.

Kayıp Veri Türleri

- **MAR** türünde veri gerçekte rastlantısal olarak kayıp değildir, **veri kaybı veri setindeki değişkenlerden bazılarına bağlıdır.**
- MAR türünde bir veri noktasının kayıp olma eğilimi kayıp veriyle ilişkili değildir ancak **gözlenen verinin bir kısmıyla İlişkilidir.**
- Rastlantısal olarak kayıp değerler ve gözlenen değerler arasında sistematik farklılıkların olabileceği ancak **bu farklılıkların diğer gözlenen değişkenlerle tamamen açıklanabileceği anlamındadır.**
- Bir değişkenin gözlemleri rastlantısal olarak kayıpsa, şartlı değişkenler kontrol edilebilirse , rastlantısal küme elde edilebilir; **kayıp ve gözlenen değerler kontrol altına alınan gruplarda benzer dağılımlara sahip olacaklardır.**

Kayıp Veri Türleri

- Büyük bir veri setinde, **verinin %5'i veya daha azı rastlantısal olarak kayıpsa çok ciddi problemlerle karşılaşılmaz** ve kayıp veri ile ilgili problemleri çözmek için kullanılan herhangi bir yöntem benzer sonuçlar verir.
- Halbuki küçük veya orta büyüklükteki bir veri setinde çok sayıda veri kaybı varsa ciddi problemler ortaya çıkabilir.
- Eldeki bilgiden yararlanarak kayıp verideki örüntüler test edilebilir.
  - Örneğin, **kayıp verinin bulunduğu değişkene göre eksik değerlere sahip bireyler ve tam değerlere sahip bireylerden iki grup oluşturulabilir. Sonra analizde bu değişkenle ilgili olabilecek diğer değişkenlerde t testi ile iki grup arasındaki ortalama farklara bakılabilir.**

Kayıp Veri

- Kayıp veriyi incelemek ve kayıp veri ile baş etmek konusunda birden fazla paket mevcuttur. Bu paketler arasında
  - VIM
  - missMethods
  - Amelia
  - naniar paketi sayılabilir.

Kayıp Veri

- İlk örnekler **naniar** üzerinden gösterilmektedir.
  - herhangi bir eksik veri olup olmadığının kontrolu

```
library(naniar)
any_na(screen)
```

```
## [1] TRUE
```

- toplam kaç eksik veri var?

```
n_miss(screen)
```

```
## [1] 27
```

- eksik veri oranı ne?

```
prop_miss(screen)
```

```
## [1] 0.008294931
```

Kayıp Veri

- eksik veriler hangi sutunlarda

```
screen %>% is.na() %>% colSums()
```

```
##    subno  timedrs  attdrug atthouse   income  mstatus     race
##        0        0        0        1       26        0        0
```

- eksik veri tablosu, frekans ve oran

```
miss_var_summary(screen)
```

```
## # A tibble: 7 × 3
##   variable n_miss pct_miss
##   <chr>     <int>    <dbl>
## 1 income       26    5.59
## 2 atthouse      1    0.215
## 3 subno         0    0
## 4 timedrs       0    0
## 5 attdrug       0    0
## 6 mstatus       0    0
## 7 race          0    0
```

Kayıp Veri

- değişkenlere göre eksik veri tablosu

```
miss_var_table(screen)
```

```
## # A tibble: 3 × 3
##   n_miss_in_var n_vars pct_vars
##           <int>  <int>    <dbl>
## 1             0      5     71.4
## 2             1      1     14.3
## 3            26      1     14.3
```

Kayıp Veri

- Hangi bireylerde/satırlarda eksik veri var

```
miss_case_summary(screen)
```

```
## # A tibble: 465 × 3
##     case n_miss pct_miss
##    <int>  <int>    <dbl>
##  1    52      1     14.3
##  2    64      1     14.3
##  3    69      1     14.3
##  4    77      1     14.3
##  5   118      1     14.3
##  6   135      1     14.3
##  7   161      1     14.3
##  8   172      1     14.3
##  9   173      1     14.3
## 10   174      1     14.3
## # … with 455 more rows
```

Kayıp Veri

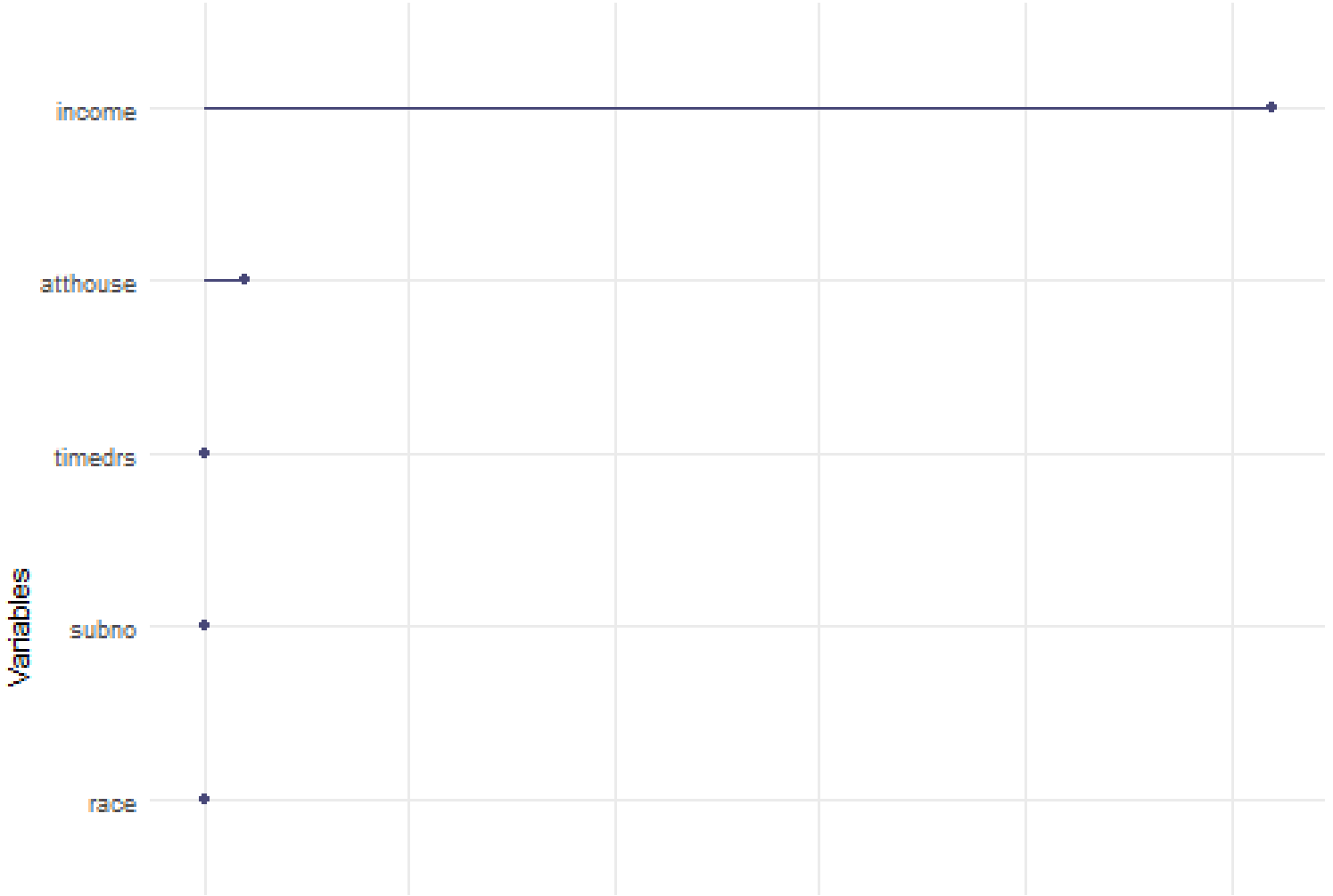
- tam ve eksik veri tablosu

```
miss_case_table(screen)
```

```
## # A tibble: 2 × 3
##   n_miss_in_case n_cases pct_cases
##            <int>   <int>     <dbl>
## 1              0     438     94.2 
## 2              1      27      5.81
```

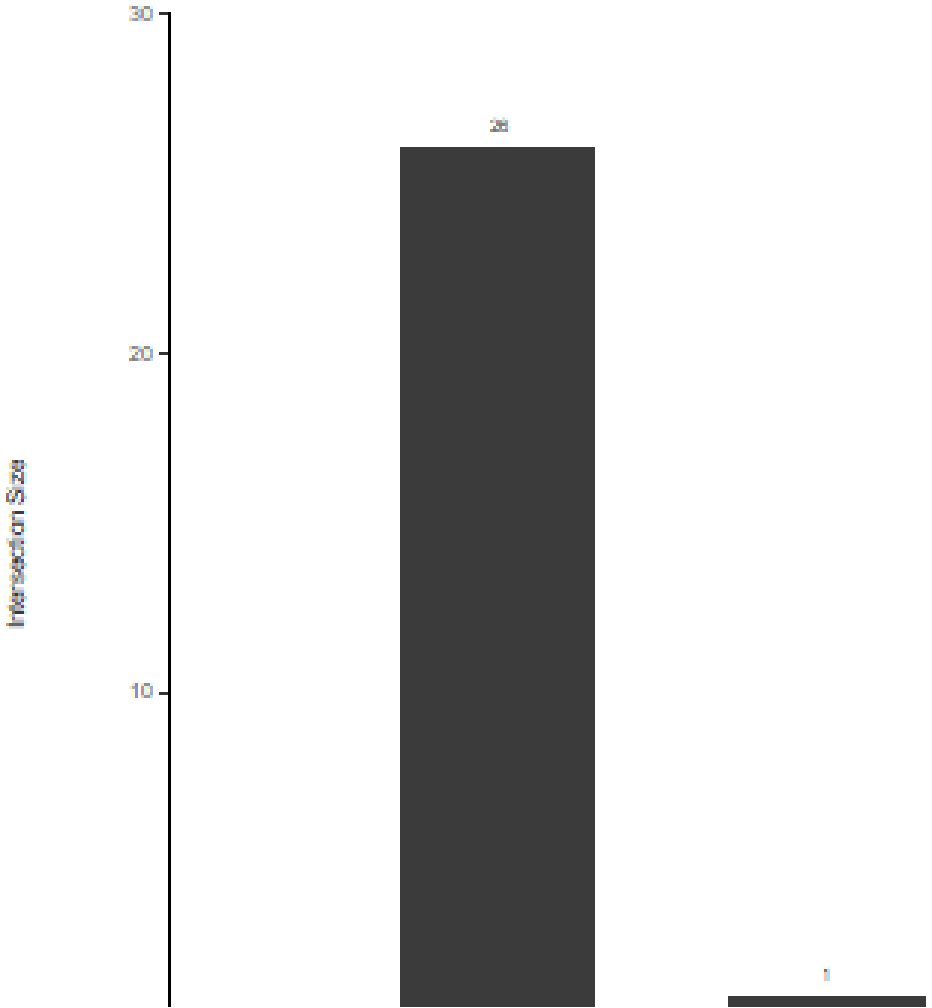
Eksik verinin görselleştirilmesi

```
gg_miss_var(screen)
```



Eksik verinin görselleştirilmesi

```
vis_miss(screen) + theme(axis.text.x = element_text(angle=80))
```



Kayıp Veri Testi

- Veri kaybının diğer değişkenlerle ilişkili olup olmadığının finalfit paketi ile incelenmesi

```
# değişkeni kopyala
screen2 <- screen
screen2$income_m <- screen2$income

library(finalfit)
explanatory=c("timedrs", "attdrug", "atthouse")
dependent="income_m"
screen2 %>%
  missing_compare(dependent, explanatory) %>%
    knitr::kable(row.names=FALSE,
        align = c("l", "l", "r", "r", "r"),
        caption = "Eksik veriye sahip olan
        ve olmayan değişkenlerin
        ortalama karşılaştırması")
```

Eksik veriye sahip olan ve olmayan değişkenlerin ortalama karşılaştırması

Missing data analysis: Income		Not missing	Missing	p
Visits to health professionals	Mean (SD)	7.9 (11.1)	7.6 (7.4)	0.891
Attitudes toward medication	Mean (SD)	7.7 (1.2)	7.9 (1.0)	0.368
Attitudes toward housework	Mean (SD)	23.5 (4.5)	23.7 (4.2)	0.860

finalfit paketi

```
screen2 <- screen %>% dplyr::mutate(mstatus =   case_when(mstatus ==1 ~ "not married",   mstatus ==2
screen2 %>%   ff_glimpse()
```

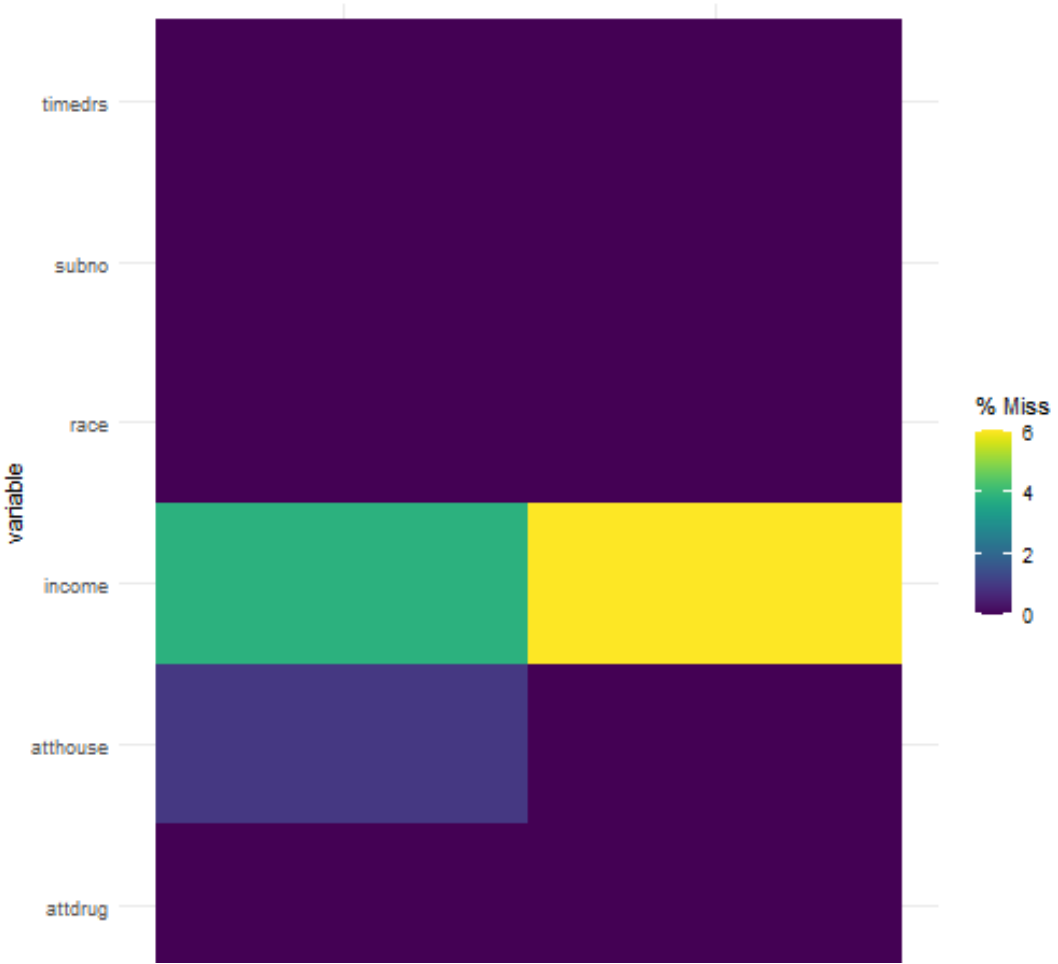
` Warning:</code>fct_explicit_na()<code>was deprecated in forcats 1.0.0. ℹ Please use</code>fct_na_value_to_level()
instead. ℹ The deprecated feature was likely used in the finalfit package. Please report the issue at <

8;;https://github.com/ewenharrison/finalfit/issues https://githu

$Continuous label var_type n missing_n subno Subject number 465 0 timedrs Visits to health professionals 465 0 attdrug Attitudes toward medication 465 0 atthouse Attitudes toward housework 464 1 income Income 439 26 race Ethnic group membership 465 0 missing_percent mean sd min quartile_25 median quartile_75 subno 0.0 317.4 194.2 1.0 137.0 314.0 483.0 timedrs 0.0 7.9 10.9 0.0 2.0 4.0 10.0 attdrug 0.0 7.7 1.2 5.0 7.0 8.0 9.0 atthouse 0.2 23.5 4.5 2.0 21.0 24.0 27.0 income 5.6 4.2 2.4 1.0 2.5 4.0 6.0 race 0.0 1.1 0.3 1.0 1.0

Bir değişkenin kategorilerinde inceleme

```
screen <- screen %>%   mutate(mstatus = as.factor(mstatus))
gg_miss_fct(screen, fct = mstatus)
```



MCAR test

```
library(naniar)
mcar_test(data=screen[,2:5])
```

```
## # A tibble: 1 × 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      3.29     6   0.772                3
```

- Little'ın MCAR testine ilişkin p değerinin . 773 olduğu görülmektedir.
- Böylece kayıp verinin MCAR olduğu sonucuna varılabilir.

Kayıp veri ile başetme

Liste bazında silme işlemi **na.omit** ve **complete.cases** fonksiyonları ile sağlanabilir.

```
na.omit(screen)
```

```
# A tibble: 438 × 7
   subno timedrs attdrug atthouse income mstatus  race
   <dbl>   <dbl>   <dbl>    <dbl>  <dbl> <fct>   <dbl>
 1     1       1       8       27      5 2          11
 2     2       3       7       20      6 2          12
 3     3       0       8       23      3 2          13
 4     4      13       9       28      8 2          14
 5     5      15       7       24      1 2          15
 6     6       3       8       25      4 2          16
 7     7       2       7       30      6 2          17
 8     8       0       7       24      6 2          18
 9     9       7       7       20      2 2          19
10    10       4       8       30      8 1          10
# … with 428 more rows
```

```
screen[complete.cases(screen),]
```

```
# A tibble: 438 × 7
   subno timedrs attdrug atthouse income mstatus  rac
   <dbl>   <dbl>   <dbl>    <dbl>  <dbl> <fct>   <dbl
 1     1       1       8       27      5 2
 2     2       3       7       20      6 2
 3     3       0       8       23      3 2
 4     4      13       9       28      8 2
 5     5      15       7       24      1 2
 6     6       3       8       25      4 2
 7     7       2       7       30      6 2
 8     8       0       7       24      6 2
 9     9       7       7       20      2 2
10    10       4       8       30      8 1
# … with 428 more rows
```

Ortalama atama

- Tek bir değişkene ortalama atama

```
df = data.frame(x = 1:20, y = c(1:10,rep(NA,10)))
df$y[is.na(df$y)] = mean(df$y, na.rm=TRUE)
```

Ortalama atama

- Tek bir değişkene ortalama atama

```
screen3 <- screen
screen3$income[is.na(screen3$income)]<- mean(screen3$income, na.rm=TRUE)
summary(screen3$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    4.00    4.21    6.00   10.00
```

Veri setindeki kayıp veriler

- **atthouse** değişkeninde bir kayıp değer bulunmaktadır ve **liste bazında silme yöntemi** ile veri setinden çıkarılmıştır.

- Veri setinde **income** değişkeni 26 kayıp değere sahiptir ve bu sayı örneklemin %5'inden fazladır. Eğer bu değişken araştırma açısından öneme sahip değilse, veri setinden çıkarılabilir, aksi halde kayıp verinin tahmin edilmesi yöntemlerinden biri kullanılabilir.

- **income** değişkenindeki kayıp değerler için kayıp verinin tahmin edilmesi yöntemlerinden ortalamanın yerleştirilmesi kullanılarak kayıp değer **yerine değişkenin ortalama değeri (4.21 değeri) yerleştirilmiştir**.

Ortalama atama

- Bu işlem birden farklı şekilde yapılabilir. Her bir sütunda eksik veriyi ortalama ile tamamlama

```
screen4 <- screen[,2:5]
for(i in 1:ncol(screen4)) {
screen4[ , i][is.na(screen4[ , i])] <- mean(screen4[ , i], na.rm = TRUE)
}
```

if_else() ile

```
# df = transform(df, y = ifelse(is.na(y), mean(y, na.rm=TRUE), y))
screen5 <- screen

screen5 = transform(screen5, income = ifelse(is.na(income), mean(income, na.rm=TRUE), income))
summary(screen5)
```

```
##      subno          timedrs          attdrug          atthouse
##  Min.   :  1.0   Min.   : 0.000   Min.   : 5.000   Min.   : 2.00
##  1st Qu.:137.0   1st Qu.: 2.000   1st Qu.: 7.000   1st Qu.:21.00
##  Median :314.0   Median : 4.000   Median : 8.000   Median :24.00
##  Mean   :317.4   Mean   : 7.901   Mean   : 7.686   Mean   :23.54
##  3rd Qu.:483.0   3rd Qu.:10.000   3rd Qu.: 9.000   3rd Qu.:27.00
##  Max.   :758.0   Max.   :81.000   Max.   :10.000   Max.   :35.00
##                                                    NA's   :1
##      income       mstatus      race
##  Min.   : 1.00   1:103   Min.   :1.000
##  1st Qu.: 3.00   2:362   1st Qu.:1.000
##  Median : 4.00           Median :1.000
##  Mean   : 4.21           Mean   :1.088
##  3rd Qu.: 6.00           3rd Qu.:1.000
##  Max.   :10.00           Max.   :2.000
##
```

mutate() ile

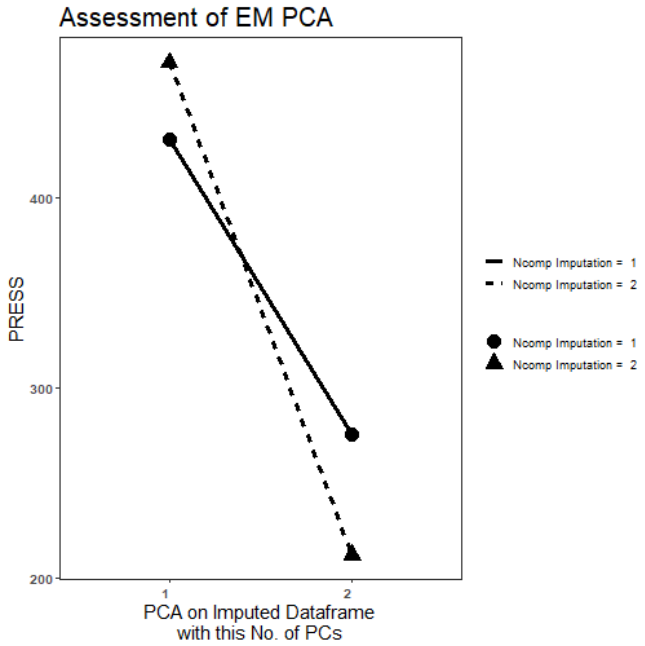
```
screen %>%
mutate(income = ifelse(is.na(income), mean(income, na.rm =TRUE), income))
```

```
## # A tibble: 465 × 7
##    subno timedrs attdrug atthouse income mstatus  race
##    <dbl>   <dbl>   <dbl>    <dbl>  <dbl> <fct>   <dbl>
##  1     1       1       8       27      5 2           1
##  2     2       3       7       20      6 2           1
##  3     3       0       8       23      3 2           1
##  4     4      13       9       28      8 2           1
##  5     5      15       7       24      1 2           1
##  6     6       3       8       25      4 2           1
##  7     7       2       7       30      6 2           1
##  8     8       0       7       24      6 2           1
##  9     9       7       7       20      2 2           1
## 10    10       4       8       30      8 1           1
## # … with 455 more rows
```

Beklenti Maksimizasyonu

- **mvdalab** paketi ile önce eksik veri olusturulup sonra eksik veri BM yöntemi ile doldurulmuştur.

```
library(mvdalab)
dat <- introNAs(iris, percent = 25)
imputeEM(dat)
```




```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width Speciessetosa
```

mvdalab paketi ile önce eksik veri oluşturulup sonra eksik veri BM yöntemi ile doldurulmuştur.

```
library(missMethods)
dat2 <- delete_MCAR(iris[,2:4], p=0.2)
dat2<-impute_EM(dat2,stochastic = FALSE)
```

Çoklu Atama

Çoklu atama için en sık kullanılan paketler **mice** ve **VIM** paketleridir.

-  mice paketi

-  Konu tekrarı

teşekkürler