title: “Group Project_Data _Processing”
author: “EMMA OO”
date: ‘2022-06-12’
output: html_document

```
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
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
```

```
df <- read.csv("/Users/emmaoo/Desktop/train.csv")
head(df)
```

```
##   Patient.Id Patient.Age Genes.in.mother.s.side Inherited.from.father
## 1  PID0x6418           2                    Yes                    No
## 2  PID0x25d5           4                    Yes                   Yes
## 3  PID0x4a82           6                    Yes                    No
## 4  PID0x4ac8          12                    Yes                    No
## 5  PID0x1bf7          11                    Yes                    No
## 6  PID0x44fe          14                    Yes                    No
##   Maternal.gene Paternal.gene Blood.cell.count..mcL. Patient.First.Name
## 1           Yes            No               4.760603            Richard
## 2            No            No               4.910669               Mike
## 3            No            No               4.893297           Kimberly
## 4           Yes            No               4.705280            Jeffery
## 5                         Yes               4.720703            Johanna
## 6           Yes            No               5.103188            Richard
##   Family.Name Father.s.name Mother.s.age Father.s.age
## 1                     Larre           NA           NA
## 2                    Brycen           NA           23
## 3                    Nashon           41           22
## 4   Hoelscher        Aayaan           21           NA
## 5    Stutzman         Suave           32           NA
## 6                  Coleston           NA           NA
##                               Institute.Name
## 1 Boston Specialty & Rehabilitation Hospital
## 2          St. Margaret’s Hospital For Women
## 3
## 4
## 5                            Carney Hospital
## 6             Massachusetts General Hospital
##                                                                        Location.of.Institute
## 1                   55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)
## 2 1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 02135\n(42.34665771451756, -71.14136122385321)
## 3                                                                                         -
## 4                   55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)
## 5       300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42.337592548462226, -71.10472284437952)
## 6                   55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)
##     Status Respiratory.Rate..breaths.min. Heart.Rate..rates.min Test.1 Test.2
## 1    Alive                 Normal (30-60)                Normal      0     NA
## 2 Deceased                      Tachypnea                Normal     NA      0
## 3    Alive                 Normal (30-60)           Tachycardia      0      0
## 4 Deceased                      Tachypnea                Normal      0      0
## 5    Alive                      Tachypnea           Tachycardia      0      0
## 6 Deceased                                               Normal      0      0
##   Test.3 Test.4 Test.5 Parental.consent Follow.up Gender Birth.asphyxia
## 1     NA      1      0              Yes      High
## 2      0      1      0              Yes      High                    No
## 3      0      1      0              Yes       Low             No record
## 4      0      1      0              Yes      High   Male  Not available
## 5      0      1      0                        Low   Male  Not available
## 6      0      1      0              Yes       Low Female  Not available
##   Autopsy.shows.birth.defect..if.applicable. Place.of.birth
## 1                             Not applicable      Institute
## 2                                       None
## 3                             Not applicable
## 4                                         No      Institute
## 5                             Not applicable      Institute
## 6                                       None      Institute
##   Folic.acid.details..peri.conceptional. H.O.serious.maternal.illness
## 1                                     No
## 2                                    Yes                          Yes
## 3                                    Yes                           No
## 4                                     No                          Yes
## 5                                     No                          Yes
## 6                                     No                           No
##   H.O.radiation.exposure..x.ray. H.O.substance.abuse
## 1                             No                  No
## 2                 Not applicable      Not applicable
## 3                            Yes
## 4                              -      Not applicable
## 5                              -      Not applicable
## 6                             No                  No
##   Assisted.conception.IVF.ART History.of.anomalies.in.previous.pregnancies
## 1                          No                                          Yes
## 2                          No                                          Yes
## 3                         Yes                                          Yes
## 4                                                                      Yes
## 5                         Yes                                           No
## 6                                                                       No
##   No..of.previous.abortion Birth.defects
## 1                       NA
## 2                       NA      Multiple
## 3                        4      Singular
## 4                        1      Singular
## 5                        4      Multiple
## 6                        0      Multiple
##   White.Blood.cell.count..thousand.per.microliter. Blood.test.result Symptom.1
## 1                                         9.857562                           1
## 2                                         5.522560            normal         1
## 3                                               NA            normal         0
## 4                                         7.919321      inconclusive         0
## 5                                         4.098210                           0
## 6                                        10.272230            normal         1
##   Symptom.2 Symptom.3 Symptom.4 Symptom.5
## 1         1         1         1         1
## 2        NA         1         1         0
## 3         1         1         1         1
## 4         0         1         0         0
## 5         0         0         0        NA
## 6         0         0         1         0
##                                Genetic.Disorder
## 1  Mitochondrial genetic inheritance disorders
## 2
## 3 Multifactorial genetic inheritance disorders
## 4  Mitochondrial genetic inheritance disorders
## 5 Multifactorial genetic inheritance disorders
## 6             Single-gene inheritance diseases
##                     Disorder.Subclass
## 1 Leber’s hereditary optic neuropathy
## 2                     Cystic fibrosis
## 3                            Diabetes
## 4                      Leigh syndrome
## 5                              Cancer
## 6                     Cystic fibrosis
```

Drop All Uninformative Columns

```
head(df)
```

```
##   Patient.Id Patient.Age Genes.in.mother.s.side Inherited.from.father
## 1  PID0x6418           2                    Yes                    No
## 2  PID0x25d5           4                    Yes                   Yes
## 3  PID0x4a82           6                    Yes                    No
## 4  PID0x4ac8          12                    Yes                    No
## 5  PID0x1bf7          11                    Yes                    No
## 6  PID0x44fe          14                    Yes                    No
##   Maternal.gene Paternal.gene Blood.cell.count..mcL. Patient.First.Name
## 1           Yes            No               4.760603            Richard
## 2            No            No               4.910669               Mike
## 3            No            No               4.893297           Kimberly
## 4           Yes            No               4.705280            Jeffery
## 5                         Yes               4.720703            Johanna
## 6           Yes            No               5.103188            Richard
##   Family.Name Father.s.name Mother.s.age Father.s.age
## 1                     Larre           NA           NA
## 2                    Brycen           NA           23
## 3                    Nashon           41           22
## 4   Hoelscher        Aayaan           21           NA
## 5    Stutzman         Suave           32           NA
## 6                  Coleston           NA           NA
##                                Institute.Name
## 1 Boston Specialty & Rehabilitation Hospital
## 2          St. Margaret’s Hospital For Women
## 3
## 4
## 5                            Carney Hospital
## 6             Massachusetts General Hospital
##                                                                        Location.of.Institute
## 1                   55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)
## 2 1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 02135\n(42.34665771451756, -71.14136122385321)
## 3                                                                                         -
## 4                   55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)
## 5       300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42.337592548462226, -71.10472284437952)
## 6                   55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)
##     Status Respiratory.Rate..breaths.min. Heart.Rate..rates.min Test.1 Test.2
## 1    Alive                 Normal (30-60)                Normal      0     NA
## 2 Deceased                      Tachypnea                Normal     NA      0
## 3    Alive                 Normal (30-60)           Tachycardia      0      0
## 4 Deceased                      Tachypnea                Normal      0      0
## 5    Alive                      Tachypnea           Tachycardia      0      0
## 6 Deceased                                               Normal      0      0
##   Test.3 Test.4 Test.5 Parental.consent Follow.up Gender Birth.asphyxia
## 1     NA      1      0              Yes      High
## 2      0      1      0              Yes      High                    No
## 3      0      1      0              Yes       Low             No record
## 4      0      1      0              Yes      High   Male  Not available
## 5      0      1      0                        Low   Male  Not available
## 6      0      1      0              Yes       Low Female  Not available
##   Autopsy.shows.birth.defect..if.applicable. Place.of.birth
## 1                             Not applicable      Institute
## 2                                       None
## 3                             Not applicable
## 4                                         No      Institute
## 5                             Not applicable      Institute
## 6                                       None      Institute
##   Folic.acid.details..peri.conceptional. H.O.serious.maternal.illness
## 1                                     No
## 2                                    Yes                          Yes
## 3                                    Yes                           No
## 4                                     No                          Yes
## 5                                     No                          Yes
## 6                                     No                           No
##   H.O.radiation.exposure..x.ray. H.O.substance.abuse
## 1                             No                  No
## 2                 Not applicable      Not applicable
## 3                            Yes
## 4                              -      Not applicable
## 5                              -      Not applicable
## 6                             No                  No
##   Assisted.conception.IVF.ART History.of.anomalies.in.previous.pregnancies
## 1                          No                                          Yes
## 2                          No                                          Yes
## 3                         Yes                                          Yes
## 4                                                                      Yes
## 5                         Yes                                           No
## 6                                                                       No
##   No..of.previous.abortion Birth.defects
## 1                       NA
## 2                       NA      Multiple
## 3                        4      Singular
## 4                        1      Singular
## 5                        4      Multiple
## 6                        0      Multiple
##   White.Blood.cell.count..thousand.per.microliter. Blood.test.result Symptom.1
## 1                                         9.857562                           1
## 2                                         5.522560            normal         1
## 3                                               NA            normal         0
## 4                                         7.919321      inconclusive         0
## 5                                         4.098210                           0
## 6                                        10.272230            normal         1
##   Symptom.2 Symptom.3 Symptom.4 Symptom.5
## 1         1         1         1         1
## 2        NA         1         1         0
## 3         1         1         1         1
## 4         0         1         0         0
## 5         0         0         0        NA
## 6         0         0         1         0
##                               Genetic.Disorder
## 1  Mitochondrial genetic inheritance disorders
## 2
## 3 Multifactorial genetic inheritance disorders
## 4  Mitochondrial genetic inheritance disorders
## 5 Multifactorial genetic inheritance disorders
## 6             Single-gene inheritance diseases
##                     Disorder.Subclass
## 1 Leber’s hereditary optic neuropathy
## 2                     Cystic fibrosis
## 3                            Diabetes
## 4                      Leigh syndrome
## 5                              Cancer
## 6                     Cystic fibrosis
```

```r
df <- subset(df, select = -c(Patient.Id,Patient.First.Name,Family.Name,Father.s.name,Institute.Name,
                             Location.of.Institute,Place.of.birth,Test.1,Test.2,Test.3,Test.5))
head(df,2)
```

```
##   Patient.Age Genes.in.mother.s.side Inherited.from.father Maternal.gene
## 1           2                    Yes                    No           Yes
## 2           4                    Yes                   Yes            No
##   Paternal.gene Blood.cell.count..mcL. Mother.s.age Father.s.age   Status
## 1            No               4.760603           NA           NA    Alive
## 2            No               4.910669           NA           23 Deceased
##   Respiratory.Rate..breaths.min. Heart.Rate..rates.min Test.4 Parental.consent
## 1                 Normal (30-60)                Normal      1              Yes
## 2                      Tachypnea                Normal      1              Yes
##   Follow.up Gender Birth.asphyxia Autopsy.shows.birth.defect..if.applicable.
## 1      High                                                   Not applicable
## 2      High                    No                                       None
##   Folic.acid.details..peri.conceptional. H.O.serious.maternal.illness
## 1                                     No
## 2                                    Yes                          Yes
##   H.O.radiation.exposure..x.ray. H.O.substance.abuse
## 1                             No                  No
## 2                 Not applicable      Not applicable
##   Assisted.conception.IVF.ART History.of.anomalies.in.previous.pregnancies
## 1                          No                                          Yes
## 2                          No                                          Yes
##   No..of.previous.abortion Birth.defects
## 1                       NA
## 2                       NA      Multiple
##   White.Blood.cell.count..thousand.per.microliter. Blood.test.result Symptom.1
## 1                                         9.857562                           1
## 2                                         5.522560            normal         1
##   Symptom.2 Symptom.3 Symptom.4 Symptom.5
## 1         1         1         1         1
## 2        NA         1         1         0
##                              Genetic.Disorder
## 1 Mitochondrial genetic inheritance disorders
## 2
##                      Disorder.Subclass
## 1 Leber's hereditary optic neuropathy
## 2                      Cystic fibrosis
```

```
dim(df)
```

```
## [1] 22083    34
```

Check Unique Values

```
# Check unique values from all columns except Blood.cell.count..mcL. and White.Blood.cell.count..thousa

#Extract those columns and save under subdf
subdf <- head(select(df, -Blood.cell.count..mcL.,-White.Blood.cell.count..thousand.per.microliter.,-Whi

list_unique <- lapply(subdf, unique)# List with unique values
list_unique
```

```
## $Patient.Age
## [1]  2  4  6 12 11 14
##
## $Genes.in.mother.s.side
## [1] "Yes"
##
## $Inherited.from.father
## [1] "No"  "Yes"
##
## $Maternal.gene
## [1] "Yes" "No"  ""
##
## $Paternal.gene
## [1] "No"  "Yes"
##
## $Mother.s.age
## [1] NA 41 21 32
##
## $Father.s.age
## [1] NA 23 22
##
## $Status
## [1] "Alive"    "Deceased"
##
## $Respiratory.Rate..breaths.min.
## [1] "Normal (30-60)" "Tachypnea"      ""
##
## $Heart.Rate..rates.min
## [1] "Normal"      "Tachycardia"
##
## $Test.4
## [1] 1
##
## $Parental.consent
## [1] "Yes" ""
##
## $Follow.up
## [1] "High" "Low"
##
## $Gender
## [1] ""       "Male"   "Female"
##
## $Birth.asphyxia
## [1] ""              "No"            "No record"     "Not available"
##
## $Autopsy.shows.birth.defect..if.applicable.
## [1] "Not applicable" "None"           "No"
##
## $Folic.acid.details..peri.conceptional.
## [1] "No"  "Yes"
##
## $H.O.serious.maternal.illness
## [1] ""    "Yes" "No"
##
## $H.O.radiation.exposure..x.ray.
## [1] "No"             "Not applicable" "Yes"            "-"
##
## $H.O.substance.abuse
## [1] "No"             "Not applicable" ""
##
## $Assisted.conception.IVF.ART
## [1] "No"  "Yes" ""
##
## $History.of.anomalies.in.previous.pregnancies
## [1] "Yes" "No"
##
## $No..of.previous.abortion
## [1] NA  4  1  0
##
## $Birth.defects
## [1] ""         "Multiple" "Singular"
##
## $Blood.test.result
## [1] ""             "normal"       "inconclusive"
##
## $Symptom.1
## [1] 1 0
##
## $Symptom.2
## [1]  1 NA  0
##
## $Symptom.3
## [1] 1 0
##
## $Symptom.4
## [1] 1 0
##
## $Symptom.5
## [1]  1  0 NA
##
## $Genetic.Disorder
## [1] "Mitochondrial genetic inheritance disorders"
## [2] ""
## [3] "Multifactorial genetic inheritance disorders"
## [4] "Single-gene inheritance diseases"
##
## $Disorder.Subclass
## [1] "Leber's hereditary optic neuropathy" "Cystic fibrosis"
## [3] "Diabetes"                            "Leigh syndrome"
## [5] "Cancer"
```

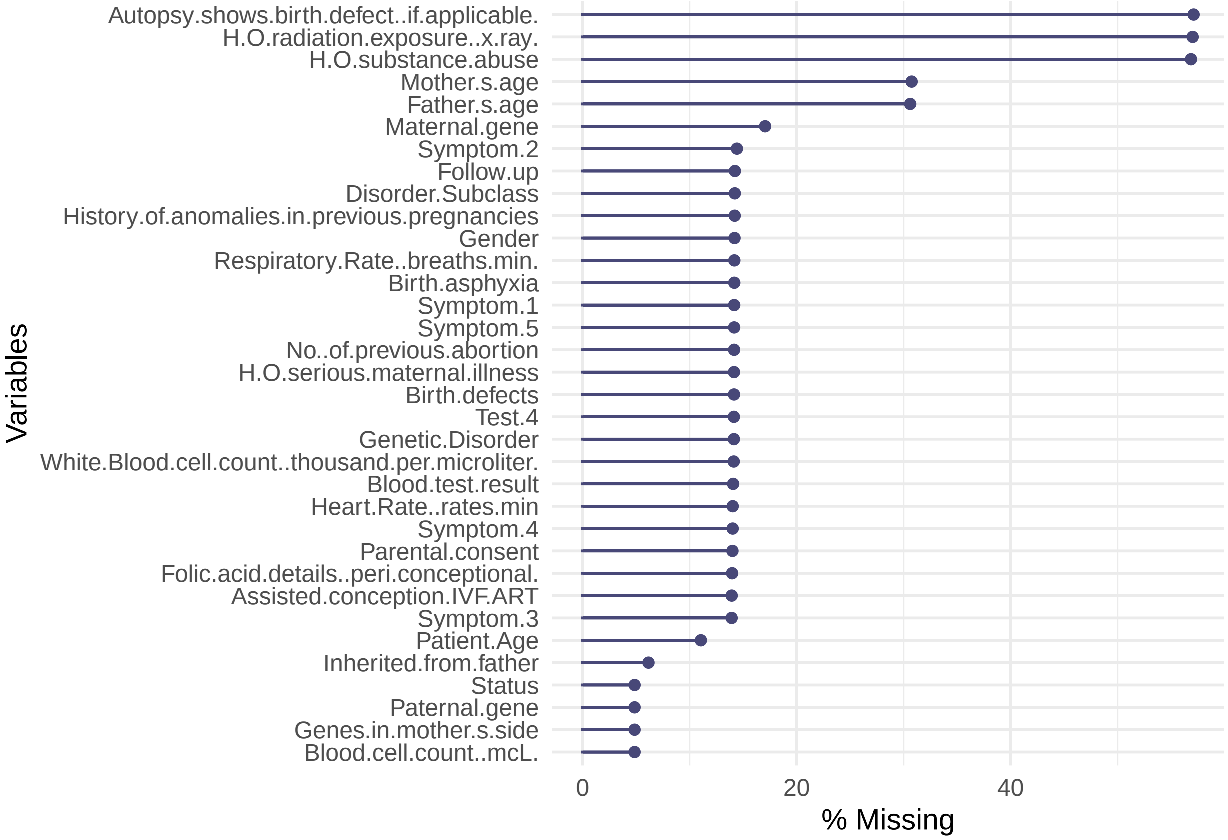
Replace ““,”_“,”Not applicable” values with NA

```
df[df == ""] <- NA
df[df == "-"] <- NA
df[df == "Not applicable"] <- NA
```

Check Missing Values

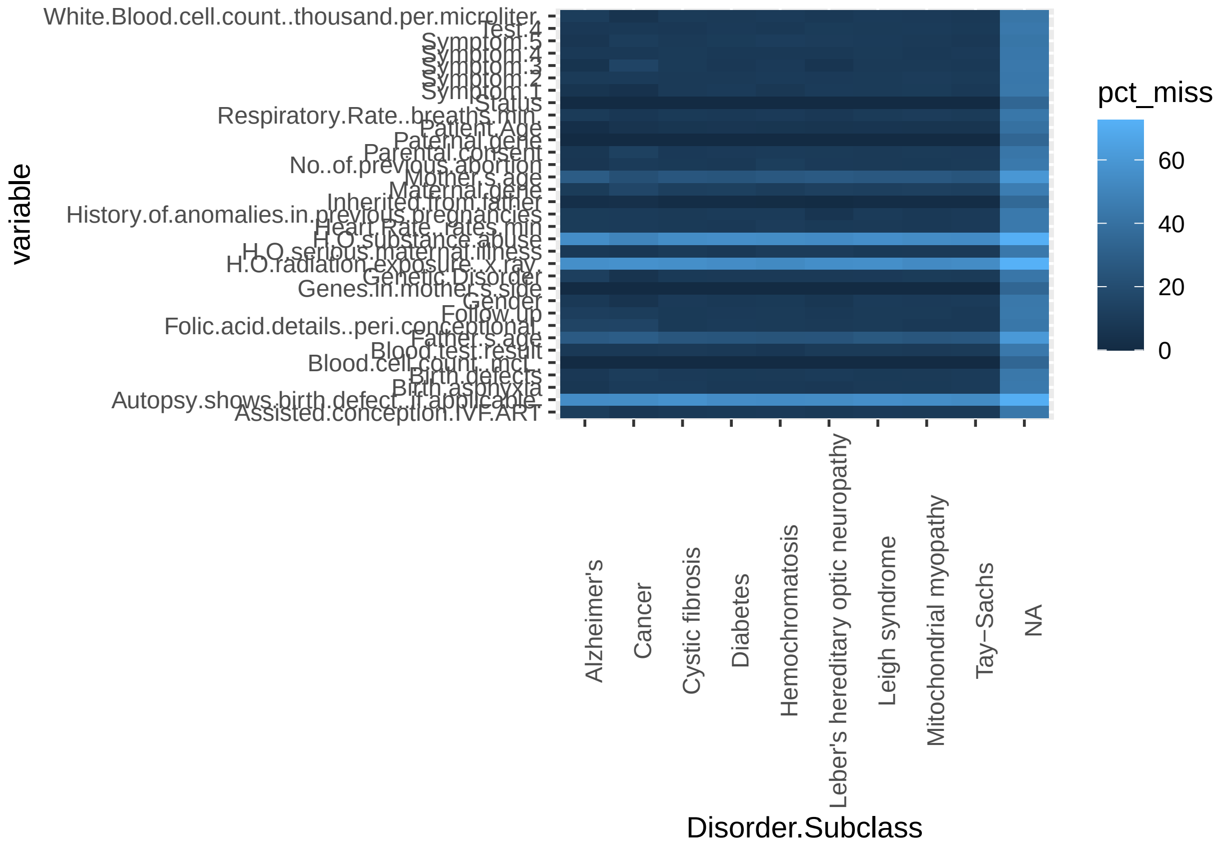
```
library(naniar)
library(dplyr)
library(caret)
gg_miss_var(df, show_pct = TRUE)
```

```
## Warning: It is deprecated to specify ‘guide = FALSE‘ to remove a guide. Please
## use ‘guide = "none"‘ instead.
```



Check if missing values were realted to the specific class

```
na_values <- df %>% group_by(Disorder.Subclass) %>% miss_var_summary()
ggplot(na_values, aes(Disorder.Subclass, variable, fill=pct_miss)) + geom_tile() +theme(axis.text.x = el
```



Data Partitioning

```
set.seed(1)
trainingrows <- createDataPartition(df$Disorder.Subclass, p = 0.80, list = FALSE)

train <- df[trainingrows,]
test <- df[trainingrows,]
```

Replace Missing Values with Median For Numerical Variables

```
library(dplyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.7     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

#for train data set
train <- train %>% mutate_if(is.numeric, funs(replace(.,is.na(.), median(., na.rm = TRUE)))) %>%
  mutate_if(is.character, funs(replace(.,is.na(.), Mode(na.omit(.)))))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
# for test data set
test <- test %>% mutate_if(is.numeric, funs(replace(.,is.na(.), median(., na.rm = TRUE)))) %>%
  mutate_if(is.character, funs(replace(.,is.na(.), Mode(na.omit(.)))))

sum(is.na(train))
```

```
## [1] 0
```

```
sum(is.na(test))
```

```
## [1] 0
```

Encoding Yes or No Columns Into Binary Columns

```
library(dplyr)
train <- train %>%mutate_if(is.character, as.factor)
test <- test %>%mutate_if(is.character,as.factor)

train <- train %>%
  mutate(across(c(Genes.in.mother.s.side,Inherited.from.father,Maternal.gene,Paternal.gene,Parental.cons
                Folic.acid.details..peri.conceptional.,H.O.serious.maternal.illness,
                  H.O.radiation.exposure..x.ray.,H.O.substance.abuse,Assisted.conception.IVF.ART,
                  History.of.anomalies.in.previous.pregnancies), ~factor(ifelse(.x == "Yes","1","0"))))

test <- test %>%
  mutate(across(c(Genes.in.mother.s.side,Inherited.from.father,Maternal.gene,Paternal.gene,Parental.cons
                  Folic.acid.details..peri.conceptional.,H.O.serious.maternal.illness,
                  H.O.radiation.exposure..x.ray.,H.O.substance.abuse,Assisted.conception.IVF.ART,
                  History.of.anomalies.in.previous.pregnancies), ~factor(ifelse(.x == "Yes","1","0"))))
```

Drop ' and - from the Disorder Subclass for Encoding Purpose

```
#For train data set
train=as.data.frame(lapply(train,gsub,pattern="'",replacement=""))
train=as.data.frame(lapply(train,gsub,pattern="-",replacement=""))
#For test data set
test=as.data.frame(lapply(test,gsub,pattern="'",replacement=""))
test=as.data.frame(lapply(test,gsub,pattern="-",replacement=""))
```

Encoding All Categorical Variables for TRAIN Data set

```
train$Status <- factor(train$Status, levels = c('Alive', 'Deceased'), labels = c(1,0))
train$Respiratory.Rate..breaths.min. <- factor(train$Respiratory.Rate..breaths.min., levels = c('Normal
train$Heart.Rate..rates.min <- factor(train$Heart.Rate..rates.min, levels = c('Normal', 'Tachycardia'),
train$Follow.up <- factor(train$Follow.up, levels = c('High', 'Low'), labels = c(1,0))
train$Gender <- factor(train$Gender, levels = c('Female', 'Male','Ambiguous'), labels = c(1,2,3))
train$Birth.asphyxia <- factor(train$Birth.asphyxia, levels = c('Yes', 'No','No record','Not available'
train$Autopsy.shows.birth.defect..if.applicable. <- factor(train$Autopsy.shows.birth.defect..if.applica

train$Birth.defects <- factor(train$Birth.defects, levels = c('Singular', 'Multiple'), labels = c(1,2))
train$Blood.test.result <- factor(train$Blood.test.result, levels = c('normal', 'slightly abnormal','ab

train$Genetic.Disorder <- factor(train$Genetic.Disorder, levels = c('Mitochondrial genetic inheritance
```

Encoding All Categorical Variables for TEST Data set

```
test$Status <- factor(test$Status, levels = c('Alive', 'Deceased'), labels = c(1,0))
test$Respiratory.Rate..breaths.min. <- factor(test$Respiratory.Rate..breaths.min., levels = c('Normal (
test$Heart.Rate..rates.min <- factor(test$Heart.Rate..rates.min, levels = c('Normal', 'Tachycardia'), l
test$Follow.up <- factor(test$Follow.up, levels = c('High', 'Low'), labels = c(1,0))
test$Gender <- factor(test$Gender, levels = c('Female', 'Male','Ambiguous'), labels = c(1,2,3))
test$Birth.asphyxia <- factor(test$Birth.asphyxia, levels = c('Yes', 'No','No record','Not available'),
test$Autopsy.shows.birth.defect..if.applicable. <- factor(test$Autopsy.shows.birth.defect..if.applicabl

test$Birth.defects <- factor(test$Birth.defects, levels = c('Singular', 'Multiple'), labels = c(1,2))
test$Blood.test.result <- factor(test$Blood.test.result, levels = c('normal', 'slightly abnormal','abno
```

```r
test$Genetic.Disorder <- factor(test$Genetic.Disorder, levels = c('Mitochondrial genetic inheritance di

head(test)
```

```
##   Patient.Age Genes.in.mother.s.side Inherited.from.father Maternal.gene
## 1           2                      1                     0             1
## 2           4                      1                     1             0
## 3           6                      1                     0             0
## 4          12                      1                     0             1
## 5          11                      1                     0             1
## 6           3                      1                     0             1
##   Paternal.gene Blood.cell.count..mcL. Mother.s.age Father.s.age Status
## 1             0            4.760603086           35           42      1
## 2             0             4.91066906           35           23      0
## 3             0            4.893297428           41           22      1
## 4             0            4.705280392           21           42      0
## 5             1            4.720702714           32           42      1
## 6             1             4.90107965           35           63      1
##   Respiratory.Rate..breaths.min. Heart.Rate..rates.min Test.4 Parental.consent
## 1                              0                     0      1                1
## 2                              1                     0      1                1
## 3                              0                     1      1                1
## 4                              1                     0      1                1
## 5                              1                     1      1                1
## 6                              0                     0      1                1
##   Follow.up Gender Birth.asphyxia Autopsy.shows.birth.defect..if.applicable.
## 1         1      3              1                                          1
## 2         1      3              0                                          0
## 3         0      3              2                                          1
## 4         1      2              2                                          0
## 5         0      2              2                                          1
## 6         0      2              2                                          1
##   Folic.acid.details..peri.conceptional. H.O.serious.maternal.illness
## 1                                      0                            1
## 2                                      1                            1
## 3                                      1                            0
## 4                                      0                            1
## 5                                      0                            1
## 6                                      1                            1
##   H.O.radiation.exposure..x.ray. H.O.substance.abuse
## 1                              0                   0
## 2                              1                   0
## 3                              1                   0
## 4                              1                   0
## 5                              1                   0
## 6                              0                   0
##   Assisted.conception.IVF.ART History.of.anomalies.in.previous.pregnancies
## 1                           0                                            1
## 2                           0                                            1
## 3                           1                                            1
## 4                           1                                            1
## 5                           1                                            0
## 6                           1                                            0
##   No..of.previous.abortion Birth.defects
## 1                        2             1
## 2                        2             2
## 3                        4             1
## 4                        1             1
## 5                        4             2
## 6                        3             2
##   White.Blood.cell.count..thousand.per.microliter. Blood.test.result Symptom.1
## 1                                      9.857562482                 1         1
## 2                                      5.522559926                 0         1
## 3                                      7.445972909                 0         0
## 4                                      7.919320981                 3         0
## 5                                      4.098210272                 1         0
## 6                                      6.825974324                 0         0
##   Symptom.2 Symptom.3 Symptom.4 Symptom.5 Genetic.Disorder
## 1         1         1         1         1                1
## 2         1         1         1         0                1
## 3         1         1         1         1                2
## 4         0         1         0         0                1
## 5         0         0         0         0                2
## 6         0         0         0         0                3
##                    Disorder.Subclass
## 1 Lebers hereditary optic neuropathy
## 2                    Cystic fibrosis
## 3                           Diabetes
## 4                     Leigh syndrome
## 5                             Cancer
## 6                           TaySachs
```

Splitting numerical and categorical predictors for visualization purpose

```
library(dplyr)
num_df <- select_if(train, is.numeric)  # Subset numeric columns with dplyr
cat_df <- select_if(train,is.character)
head(num_df)
```

```
## data frame with 0 columns and 6 rows
```

```
head(cat_df)
```

```
##   Patient.Age Genes.in.mother.s.side Inherited.from.father Maternal.gene
## 1           2                      1                     0             1
## 2           4                      1                     1             0
## 3           6                      1                     0             0
## 4          12                      1                     0             1
## 5          11                      1                     0             1
## 6           3                      1                     0             1
##   Paternal.gene Blood.cell.count..mcL. Mother.s.age Father.s.age Test.4
## 1             0            4.760603086           35           42      1
## 2             0             4.91066906           35           23      1
## 3             0            4.893297428           41           22      1
## 4             0            4.705280392           21           42      1
## 5             1            4.720702714           32           42      1
## 6             1             4.90107965           35           63      1
##   Parental.consent Folic.acid.details..peri.conceptional.
## 1                1                                      0
## 2                1                                      1
## 3                1                                      1
## 4                1                                      0
## 5                1                                      0
## 6                1                                      1
##   H.O.serious.maternal.illness H.O.radiation.exposure..x.ray.
## 1                            1                              0
## 2                            1                              1
## 3                            0                              1
## 4                            1                              1
## 5                            1                              1
## 6                            1                              0
##   H.O.substance.abuse Assisted.conception.IVF.ART
## 1                   0                           0
## 2                   0                           0
## 3                   0                           1
## 4                   0                           1
## 5                   0                           1
## 6                   0                           1
##   History.of.anomalies.in.previous.pregnancies No..of.previous.abortion
## 1                                            1                        2
## 2                                            1                        2
## 3                                            1                        4
## 4                                            1                        1
## 5                                            0                        4
## 6                                            0                        3
##   White.Blood.cell.count..thousand.per.microliter. Symptom.1 Symptom.2
## 1                                      9.857562482         1         1
## 2                                      5.522559926         1         1
## 3                                      7.445972909         0         1
## 4                                      7.919320981         0         0
## 5                                      4.098210272         0         0
## 6                                      6.825974324         0         0
##   Symptom.3 Symptom.4 Symptom.5                  Disorder.Subclass
## 1         1         1         1 Lebers hereditary optic neuropathy
## 2         1         1         0                    Cystic fibrosis
## 3         1         1         1                           Diabetes
## 4         1         0         0                     Leigh syndrome
## 5         0         0         0                             Cancer
## 6         0         0         0                           TaySachs
```

Splitting train data into predictors and outcome

```
trainX <- train[,-34]
trainy <- train$Disorder.Subclass
```

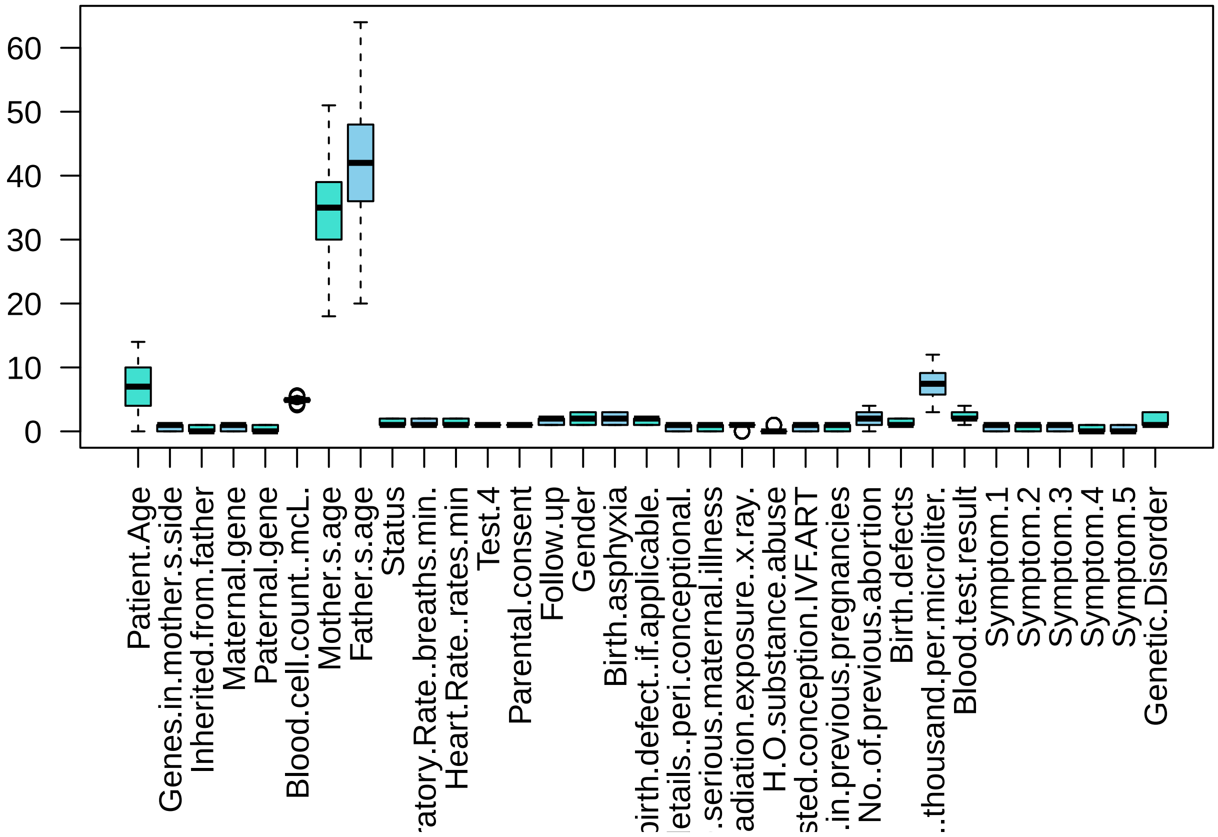
```
testX <- test[,-34]
testy <- test$Disorder.Subclass
```

Change all char and factor into numeric variables

```
trainX <- trainX %>% mutate_if(is.character, as.numeric)
trainX <- trainX %>% mutate_if(is.factor, as.numeric)
```

#Boxplot

```
par(mar=c(10,2,1,1))
boxplot(trainX, las=2, col = c("turquoise","skyblue"))
```



Countplot

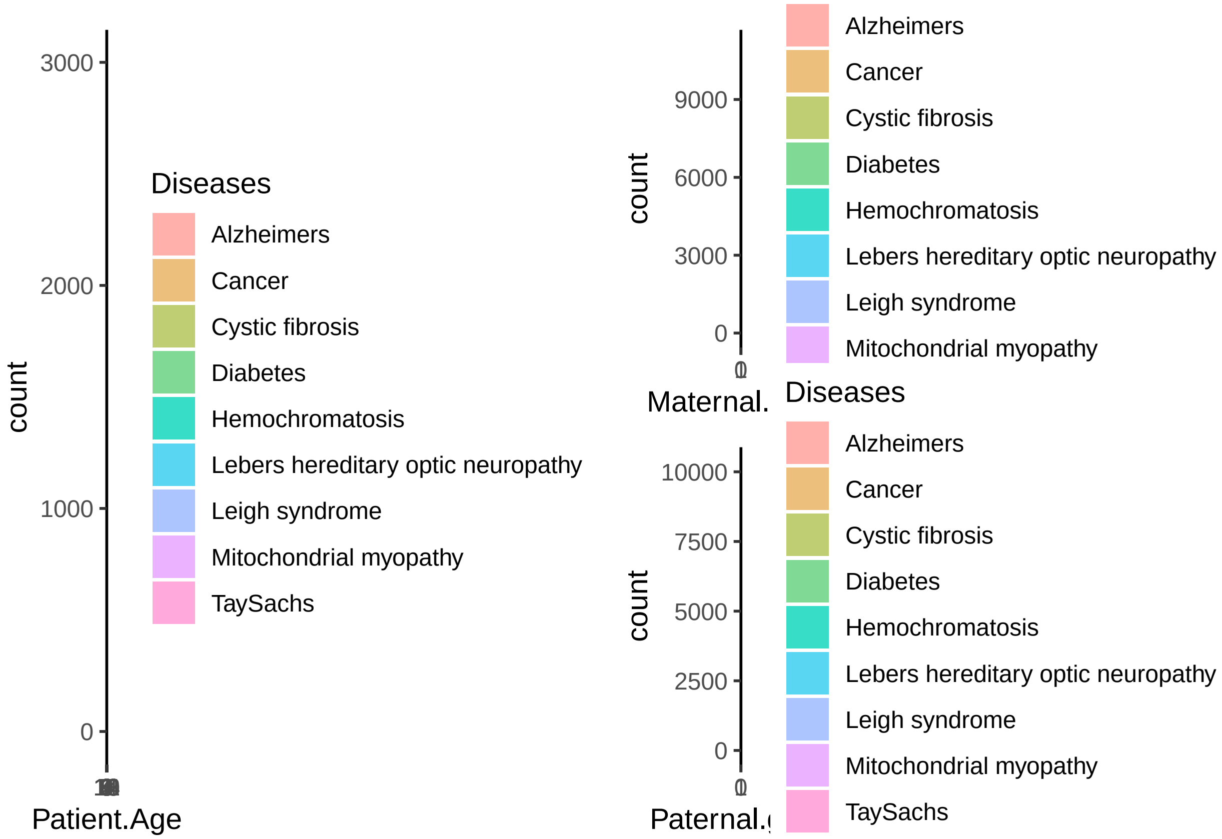
```
library(ggplot2)
library(patchwork)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:patchwork':
##
##     align_plots
```

```
Diseases <- trainy
p1 <- ggplot(train, aes(x = Patient.Age, fill = Diseases)) +geom_bar() + theme_classic()+scale_fill_hue
p2 <- ggplot(train, aes(x = Maternal.gene, fill = Diseases)) +geom_bar() +theme_classic()+ scale_fill_hu
p3 <- ggplot(train, aes(x = Paternal.gene, fill = Diseases)) +geom_bar() +theme_classic()+ scale_fill_hu

p1+p2/p3
```



Check highly correlated predictors

```
corr <- cor(trainX)
```

```
## Warning in cor(trainX): the standard deviation is zero
```

```
highcor <- findCorrelation(corr, 0.70)
colnames(train)[highcor]
```

```
## character(0)
```

Check Near Zero Variance Predictors and Dropping them

```
trainX <- trainX[,-nearZeroVar(trainX)]
testX <- testX[,-nearZeroVar(testX)]
dim(trainX)
```

```
## [1] 17670    31
```

```
dim(testX)
```

```
## [1] 17670    31
```

Skewness

```
library(moments)
skewness(trainX)
```

```
##                                      Patient.Age
##                                      0.008296585
##                           Genes.in.mother.s.side
##                                     -0.482007289
##                            Inherited.from.father
##                                      0.525953319
##                                    Maternal.gene
##                                     -0.537489794
##                                    Paternal.gene
##                                      0.351426045
##                           Blood.cell.count..mcL.
##                                     -0.001706884
##                                     Mother.s.age
##                                     -0.058629347
##                                     Father.s.age
##                                     -0.012958648
##                                           Status
##                                      0.117007879
##                   Respiratory.Rate..breaths.min.
##                                      0.304780085
##                            Heart.Rate..rates.min
##                                      0.331572929
##                                        Follow.up
##                                     -0.312753559
##                                           Gender 
##                                     -0.284802739 
##                                   Birth.asphyxia 
##                                     -0.141891353 
##       Autopsy.shows.birth.defect..if.applicable. 
##                                     -0.972658688 
##           Folic.acid.details..peri.conceptional. 
##                                     -0.305951756 
##                     H.O.serious.maternal.illness 
##                                     -0.282108790 
##                   H.O.radiation.exposure..x.ray. 
##                                     -1.387889615 
##                              H.O.substance.abuse 
##                                      1.390750080 
##                      Assisted.conception.IVF.ART 
##                                     -0.282808327 
##     History.of.anomalies.in.previous.pregnancies 
##                                     -0.316980688 
##                         No..of.previous.abortion 
##                                     -0.001473218 
##                                    Birth.defects 
##                                      0.297288597 
## White.Blood.cell.count..thousand.per.microliter. 
##                                      0.023043841 
##                                Blood.test.result 
##                                      0.201314025 
##                                        Symptom.1 
##                                     -0.622850779 
##                                        Symptom.2 
##                                     -0.475116772 
##                                        Symptom.3 
##                                     -0.427073036 
##                                        Symptom.4 
##                                      0.294716238 
##                                        Symptom.5 
##                                      0.425379260 
##                                 Genetic.Disorder 
##                                      0.511253247
```