Covid-19 effect on Liver Cancer (Handling Missing records)

Oghenekeno Eribewe

2023-11-11

Brief overview of the data set

Loading the data set gotten from https://www.kaggle.com/datasets/fedesoriano/covid19-effect-on-liver-cancer-prediction-dataset with strings loaded as factors and header set to true as the data comes with headers.

```
df = read.csv("covid-liver.csv", header = T, stringsAsFactors = T)
```

Next, we take a brief view and stats of the data to analyse the missing values and proceed with what ever solution we can come up with to sort those missing values.

```
#View(df)
head(df, 5)
```

```
##   Cancer        Year Month Bleed Mode_Presentation Age Gender Etiology
## 1      Y Prepandemic     1     N      Surveillance  68      M    NAFLD
## 2      Y Prepandemic     1     N      Surveillance  70      M     ARLD
## 3      Y Prepandemic     1     N      Surveillance  64      M     ARLD
## 4      Y Prepandemic     1     N        Incidental  73      M     ARLD
## 5      Y Prepandemic     1     N        Incidental  66      F     ARLD
##   Cirrhosis Size HCC_TNM_Stage HCC_BCLC_Stage ICC_TNM_Stage
Treatment_grps
## 1         Y   22            II              A          <NA>
Ablation
## 2         Y   40             I              D          <NA> Supportive
care
## 3         Y   52            IV              B          <NA>
Medical
## 4         Y   80            IV              C          <NA> Supportive
care
## 5         Y   60             I              0          <NA> Supportive
care
##   Survival_fromMDM Alive_Dead Type_of_incidental_finding
Surveillance_programme
## 1            32.73      Alive                       <NA>
Y
## 2             3.03       Dead                       <NA>
Y
## 3            14.97       Dead                       <NA>
Y
## 4             1.40       Dead    Secondary care\x97acute
N
```

```
## 5            32.50      Alive    Secondary care\x97acute
N
##   Surveillance_effectiveness Mode_of_surveillance_detection
## 1                 Consistent                             US
## 2                 Consistent                             US
## 3                 Consistent                             US
## 4                       <NA>                           <NA>
## 5                       <NA>                           <NA>
##   Time_diagnosis_1st_Tx Date_incident_surveillance_scan PS
## 1                  0.47                           <NA>  0
## 2                    NA                           <NA>  2
## 3                    NA                           <NA>  0
## 4                    NA                           <NA>  2
## 5                    NA                           <NA>  0
##   Time_MDM_1st_treatment Time_decisiontotreat_1st_treatment
## 1                    0.7                                NA
## 2                     NA                                NA
## 3                     NA                                NA
## 4                     NA                                NA
## 5                     NA                                NA
##   Prev_known_cirrhosis Months_from_last_surveillance
## 1                    Y                      7.333333
## 2                    Y                      4.033333
## 3                    Y                      5.900000
## 4                    Y                            NA
## 5                    Y                            NA
```

```
summary(df)
```

```
##  Cancer           Year         Month         Bleed
Mode_Presentation
##  N:140   Pandemic   :184   Min.   : 1.000   N   :304   Incidental  :145
##  Y:310   Prepandemic:266   1st Qu.: 4.000   Y   :  6   Surveillance:104
##                            Median : 7.000   NA's:140   Symptomatic :201
##                            Mean   : 6.758
##                            3rd Qu.:10.000
##                            Max.   :12.000
##
##       Age        Gender                Etiology   Cirrhosis       Size
##  Min.   :27.00   F:115   NAFLD             :120   N   : 96   Min.   :
10.00
##  1st Qu.:65.00   M:335   ARLD              : 95   Y   :215   1st Qu.:
24.00
##  Median :72.00           No established CLD: 38   NA's:139   Median :
40.00
##  Mean   :70.37           HCV               : 24              Mean   :
53.35
##  3rd Qu.:78.00           HH                : 15              3rd Qu.:
70.50
##  Max.   :96.00           (Other)           : 19              Max.
```

```
:220.00
##                          NA's              :139              NA's   :50
##    HCC_TNM_Stage HCC_BCLC_Stage ICC_TNM_Stage         Treatment_grps
##  I        :127   0   :  2       I   :  2      Supportive care:236
##  II       : 58   A   : 65       II  : 22      TACE           : 62
##  IIIA+IIIB: 86   B   : 25       III : 20      Medical        : 51
##  IV       : 40   C   :152       IV  : 95      Ablation       : 40
##  NA's     :139   D   : 67       NA's:311      Resection      : 30
##                  NA's:139                     (Other)        : 29
##                                               NA's           :  2
##  Survival_fromMDM Alive_Dead  Type_of_incidental_finding
##  Min.   :-0.030   Alive:186   Primary care\x97acute    :  5
##  1st Qu.: 4.032   Dead :264   Secondary care\x97routine: 23
##  Median :10.785               Primary care\x97routine  : 30
##  Mean   :12.697               Secondary care\x97acute  : 66
##  3rd Qu.:21.282               NA's                     :326
##  Max.   :32.770
##
##  Surveillance_programme Surveillance_effectiveness
##  N   :190               Consistent  : 78
##  Y   :121               Inconsistent: 21
##  NA's:139               Missed      : 18
##                         NA's        :333
##
##
##
##  Mode_of_surveillance_detection Time_diagnosis_1st_Tx
##  AFP alone: 24                  Min.   :-1434.070
##  CT/MRI   :  4                  1st Qu.:    1.208
##  US       : 70                  Median :    1.915
##  NA's     :352                  Mean   :   -6.570
##                                 3rd Qu.:    3.160
##                                 Max.   :   13.570
##                                 NA's   :292
##  Date_incident_surveillance_scan       PS        Time_MDM_1st_treatment
##  N   : 12                         Min.   :0.000   Min.   :-0.870
##  Y   : 21                         1st Qu.:0.000   1st Qu.: 1.185
##  NA's:417                         Median :1.000   Median : 1.800
##                                   Mean   :1.225   Mean   : 2.386
##                                   3rd Qu.:2.000   3rd Qu.: 2.723
##                                   Max.   :4.000   Max.   :15.000
##                                   NA's   :2       NA's   :288
##  Time_decisiontotreat_1st_treatment Prev_known_cirrhosis
##  Min.   :-0.870                     N   :172
##  1st Qu.: 0.715                     Y   :273
##  Median : 1.370                     NA's:  5
##  Mean   : 1.501
##  3rd Qu.: 1.715
##  Max.   :11.900
##  NA's   :343
```

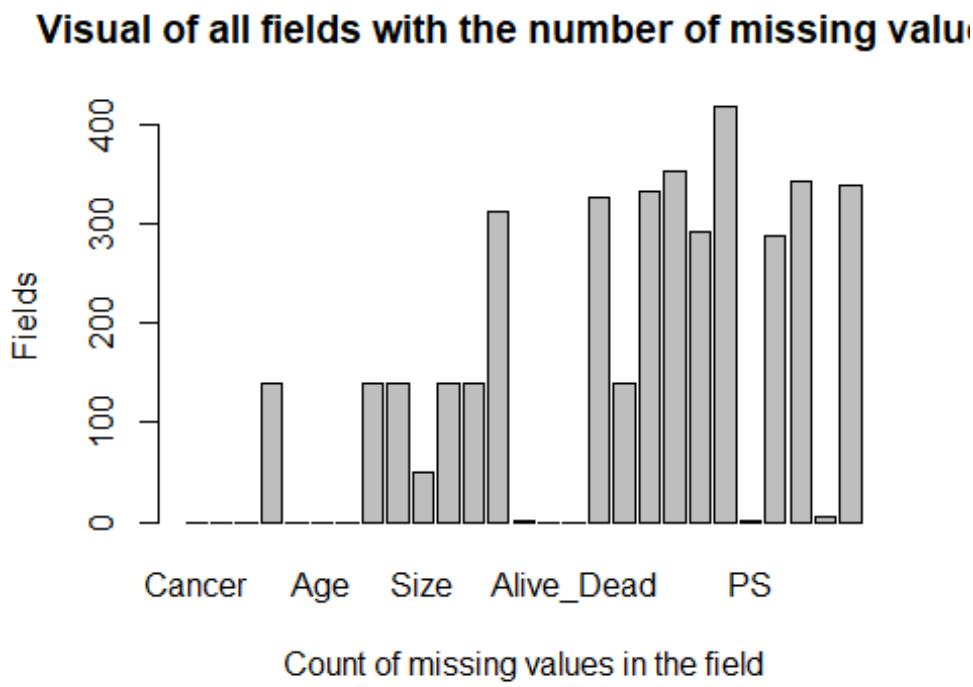
```
##  Months_from_last_surveillance
##  Min.   : 0.330
##  1st Qu.: 5.567
##  Median : 6.300
##  Mean   :10.650
##  3rd Qu.:10.600
##  Max.   :82.433
##  NA's   :338
```

It is observed that there are quite a number of fields with missing values, so we need to visualize these fields to observe the levels of these missing values in the different fields.

```
# checking the sum of NA's in all columns
# sapply(covidLiver, anyNA)
df.Na = colSums(is.na(df))
print(df.Na)
```

```
##                              Cancer                                Year
##                                   0                                   0
##                               Month                               Bleed
##                                   0                                 140
##                  Mode_Presentation                                 Age
##                                   0                                   0
##                              Gender                            Etiology
##                                   0                                 139
##                           Cirrhosis                                Size
##                                 139                                  50
##                       HCC_TNM_Stage                      HCC_BCLC_Stage
##                                 139                                 139
##                       ICC_TNM_Stage                      Treatment_grps
##                                 311                                   2
##                    Survival_fromMDM                          Alive_Dead
##                                   0                                   0
##         Type_of_incidental_finding              Surveillance_programme
##                                 326                                 139
##         Surveillance_effectiveness      Mode_of_surveillance_detection
##                                 333                                 352
##              Time_diagnosis_1st_Tx     Date_incident_surveillance_scan
##                                 292                                 417
##                                 PS              Time_MDM_1st_treatment
##                                  2                                 288
## Time_decisiontotreat_1st_treatment                Prev_known_cirrhosis
##                                 343                                   5
##      Months_from_last_surveillance
##                                 338
```

```
# visualizing the number of columns with the quantity of  missing values
barplot(df.Na, xlab = "Count of missing values in the field", ylab =
"Fields", main = "Visual of all fields with the number of missing values")
```



Quite a number of the fields have missing values consisting of more than 50% of the field and there is no way we can predict those missing values as there is not enough remaining data in the fields to make such prediction/assumption. Hence, there is need to remove these fields.

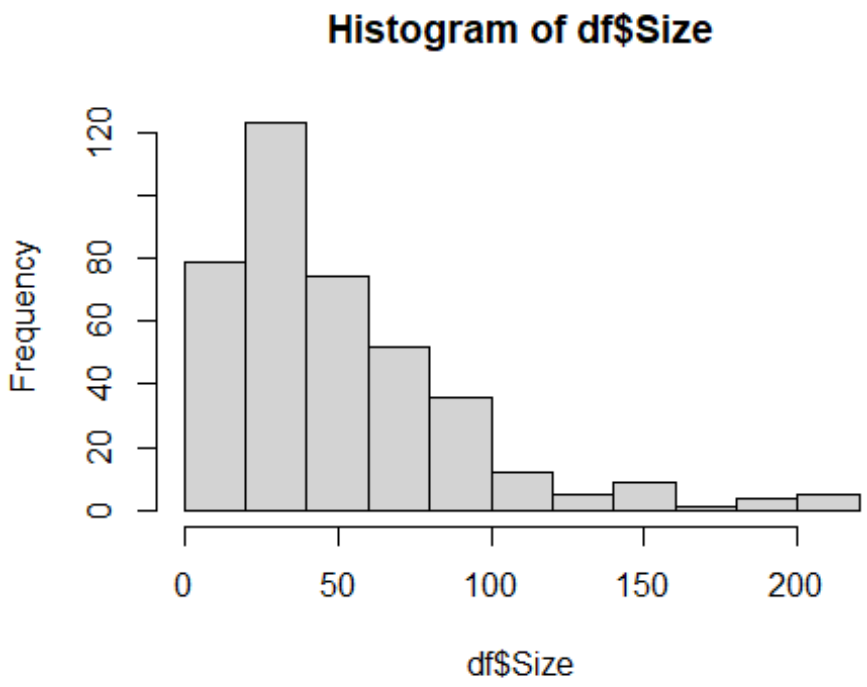
```
# removing the columns using their field numbers
df[,c(13,17,19,20,21,22,24,25,27)] = NULL
summary(df)
```

```
##  Cancer           Year         Month          Bleed
Mode_Presentation
##  N:140   Pandemic   :184   Min.   : 1.000   N   :304   Incidental  :145
##  Y:310   Prepandemic:266   1st Qu.: 4.000   Y   :  6   Surveillance:104
##                            Median : 7.000   NA's:140   Symptomatic :201
##                            Mean   : 6.758
##                            3rd Qu.:10.000
##                            Max.   :12.000
##
##       Age        Gender                Etiology   Cirrhosis       Size
##  Min.   :27.00   F:115   NAFLD             :120   N   : 96   Min.   :
10.00
##  1st Qu.:65.00   M:335   ARLD              : 95   Y   :215   1st Qu.:
24.00
##  Median :72.00           No established CLD: 38   NA's:139   Median :
40.00
##  Mean   :70.37           HCV               : 24              Mean   :
53.35
```

```
##  3rd Qu.:78.00           HH                : 15             3rd Qu.:
70.50
##  Max.   :96.00           (Other)           : 19             Max.
:220.00
##                          NA's              :139             NA's   :50
##    HCC_TNM_Stage HCC_BCLC_Stage         Treatment_grps Survival_fromMDM
##  I        :127   0   :  2       Supportive care:236    Min.   :-0.030
##  II       : 58   A   : 65       TACE           : 62    1st Qu.: 4.032
##  IIIA+IIIB: 86   B   : 25       Medical        : 51    Median :10.785
##  IV       : 40   C   :152       Ablation       : 40    Mean   :12.697
##  NA's     :139   D   : 67       Resection      : 30    3rd Qu.:21.282
##                  NA's:139       (Other)        : 29    Max.   :32.770
##                                 NA's           :  2
##  Alive_Dead  Surveillance_programme       PS        Prev_known_cirrhosis
##  Alive:186   N   :190              Min.   :0.000   N   :172
##  Dead :264   Y   :121              1st Qu.:0.000   Y   :273
##              NA's:139              Median :1.000   NA's:  5
##                                    Mean   :1.225
##                                    3rd Qu.:2.000
##                                    Max.   :4.000
##                                    NA's   :2
```

Now, we can fix up the remaining fields. We start by fixing the fields with very small amount of missing values as it won't require too much technicalities Starting from the size, treatment_grps, PS, and Prev_Known_Cirrhosis fields.

```
# quick visual of the distribution of the values in the field
hist(x = df$Size, freq = TRUE)
```



The distribution is positively skewed, and so the best option would be to impute the median as it would be a more representative value than the mean.

```
df$Size = impute((df$Size), median)
summary(df$Size)

##
##  50 values imputed to 40

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   25.00   40.00   51.87   66.50  220.00
```

Next, we impute the mode for the treatment_grps as it only has 2 missing values and we do not want to drop the rows

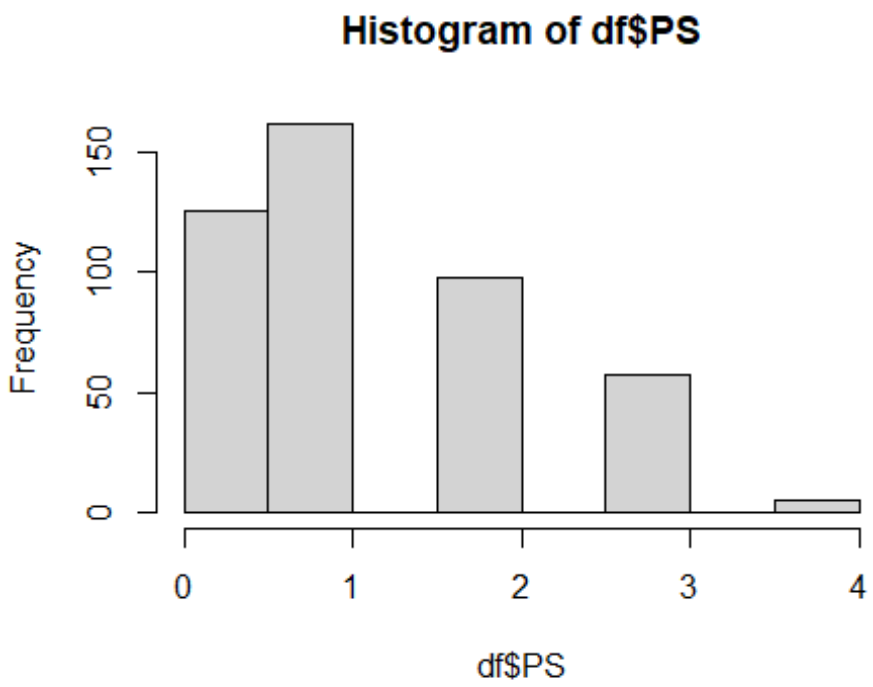
```
# next we handle the field treatment_grps
df$Treatment_grps = impute((df$Treatment_grps), mode)
summary(df$Treatment_grps)

##
##  2 values imputed to Supportive care

##        Ablation         Medical            OLTx       Resection
SIRT
##              40              51               7              30
22
## Supportive care            TACE
##             238              62
```

Next, we visualise the distribution for the PS as it only has 2 missing values

```
# quick visual of the distribution of the values in the field
hist(x = df$PS)
```




```
# the median seems ideal to impute as the distribution is awkwardly skewed
df$PS = impute((df$PS), median)
summary(df$PS)

##
##  2 values imputed to 1

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.224   2.000   4.000

summary(df)

##
##  50 values imputed to 40
##
##
##  2 values imputed to Supportive care
##
##
##  2 values imputed to 1

##  Cancer           Year         Month         Bleed
Mode_Presentation
```

```
##  N:140   Pandemic   :184   Min.   : 1.000   N   :304   Incidental  :145
##  Y:310   Prepandemic:266   1st Qu.: 4.000   Y   :  6   Surveillance:104
##                            Median : 7.000   NA's:140   Symptomatic :201
##                            Mean   : 6.758
##                            3rd Qu.:10.000
##                            Max.   :12.000
##
##       Age        Gender                Etiology   Cirrhosis       Size
##  Min.   :27.00   F:115   NAFLD             :120   N   : 96   Min.   :
10.00
##  1st Qu.:65.00   M:335   ARLD              : 95   Y   :215   1st Qu.:
25.00
##  Median :72.00           No established CLD: 38   NA's:139   Median :
40.00
##  Mean   :70.37           HCV               : 24              Mean   :
51.87
##  3rd Qu.:78.00           HH                : 15              3rd Qu.:
66.50
##  Max.   :96.00           (Other)           : 19              Max.
:220.00
##                          NA's              :139
##    HCC_TNM_Stage HCC_BCLC_Stage         Treatment_grps Survival_fromMDM
##  I        :127   0   :  2       Ablation       : 40    Min.   :-0.030
##  II       : 58   A   : 65       Medical        : 51    1st Qu.: 4.032
##  IIIA+IIIB: 86   B   : 25       OLTx           :  7    Median :10.785
##  IV       : 40   C   :152       Resection      : 30    Mean   :12.697
##  NA's     :139   D   : 67       SIRT           : 22    3rd Qu.:21.282
##                  NA's:139       Supportive care:238    Max.   :32.770
##                                 TACE           : 62
##  Alive_Dead  Surveillance_programme       PS        Prev_known_cirrhosis
##  Alive:186   N   :190               Min.   :0.000   N   :172
##  Dead :264   Y   :121               1st Qu.:0.000   Y   :273
##              NA's:139               Median :1.000   NA's:  5
##                                     Mean   :1.224
##                                     3rd Qu.:2.000
##                                     Max.   :4.000
##
```

Next, we impute the mode for the Prev_known_cirrhosis as it only has 5 missing values and we do not want to drop the rows

```
# next we handle the field Prev_known_cirrhosis
df$Prev_known_cirrhosis = impute((df$Prev_known_cirrhosis), mode)
summary(df$Prev_known_cirrhosis)
```

```
##
##  5 values imputed to Y
```

```
##   N   Y
## 172 278
```

From the mice library, we use the random forest mice function to predict the missing categorical values.

So we use the 5th cycle out of the 5 cycles of predictions made.

```
cleaned_df = complete(imputed_data, 5)
summary(cleaned_df)
```

```
##
##  2 values imputed to Supportive care
##
##
##  5 values imputed to Y
```

```
##  Cancer           Year         Month        Bleed      Mode_Presentation
##  N:140   Pandemic   :184   Min.   : 1.000   N:444   Incidental  :145
##  Y:310   Prepandemic:266   1st Qu.: 4.000   Y:  6   Surveillance:104
##                            Median : 7.000           Symptomatic :201
##                            Mean   : 6.758
##                            3rd Qu.:10.000
##                            Max.   :12.000
##
##       Age        Gender                 Etiology   Cirrhosis      Size
##  Min.   :27.00   F:115   NAFLD              :160   N:119     Min.   : 10.00
##  1st Qu.:65.00   M:335   ARLD               :153   Y:331     1st Qu.: 25.00
##  Median :72.00           No established CLD: 49             Median : 40.00
##  Mean   :70.37           HCV                : 36             Mean   : 51.87
##  3rd Qu.:78.00           HH                 : 22             3rd Qu.: 66.50
##  Max.   :96.00           PBC/AIH            : 20             Max.   :220.00
##                          (Other)            : 10
##    HCC_TNM_Stage HCC_BCLC_Stage         Treatment_grps Survival_fromMDM
##  I        :178   0:  2          Ablation       : 40    Min.   :-0.030
##  II       : 84   A: 93          Medical        : 51    1st Qu.: 4.032
##  IIIA+IIIB:125   B: 30          OLTx           :  7    Median :10.785
##  IV       : 63   C:223          Resection      : 30    Mean   :12.697
##                  D:102          SIRT           : 22    3rd Qu.:21.282
##                                 Supportive care:238    Max.   :32.770
##                                 TACE           : 62
##  Alive_Dead  Surveillance_programme       PS        Prev_known_cirrhosis
##  Alive:186   N:289                  Min.   :0.000   N:172
##  Dead :264   Y:161                  1st Qu.:0.000   Y:278
##                                     Median :1.000
##                                     Mean   :1.224
##                                     3rd Qu.:2.000
##                                     Max.   :4.000
##
```

```
df1 = cleaned_df

write.csv(df1, "clean-CovidLiver.csv", row.names = FALSE)
```

In summary, we have carried out simple data processing to handle the messy dataset to generate a clean dataset which can now be used for proper analysis and exploration.