Proyecto Final

Luis Antonio Guerrero Ibarra

```
library(magrittr)
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
```

```
library(ggplot2)
```

LOS SIGUIENTES PUNTOS SE CONSIDERARAN PARA LA EVALUACION: Buenas practicas Codigo limpio y legible Esfuerzo Resultados Comentarios y observaciones del analisis que vayan haciendo

HINT GENERAL: Si no lo sabes hacer, googlealo o revisa los scripts de las clases Mas informacion sobre el dataset utilizado: # https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

1) Descarga el csv de los datos del COVID http://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip, importa los datos en R RECOMENDACION: usar read_csv

```
covid <- read.csv("covid_dataset.csv")
```

2) Extrae una muestra aleatoria de 10k registros y asignala en una nueva variable. A partir de ahora trabaja con este dataset # HINT: usar funcion sample_n de dplyr

```
covid_aleatorio <- covid %>% sample_n(10000)
```

3) Haz un resumen estadistico del dataset y tambien muestra los tipos de datos por columna

```
summary(covid)
```

```
##  FECHA_ACTUALIZACION ID_REGISTRO           ORIGEN         SECTOR
##  Length:100000       Length:100000      Min.   :1.00   Min.   : 1.000
##  Class :character    Class :character   1st Qu.:1.00   1st Qu.: 4.000
##  Mode  :character    Mode  :character   Median :2.00   Median :12.000
##                                         Mean   :1.67   Mean   : 9.406
##                                         3rd Qu.:2.00   3rd Qu.:12.000
##                                         Max.   :2.00   Max.   :13.000
##    ENTIDAD_UM         SEXO        ENTIDAD_NAC     ENTIDAD_RES
##  Min.   : 1.00   Min.   :1.000   Min.   : 1.00   Min.   : 1.0
##  1st Qu.: 9.00   1st Qu.:1.000   1st Qu.: 9.00   1st Qu.: 9.0
##  Median :11.00   Median :1.000   Median :14.00   Median :14.0
##  Mean   :14.51   Mean   :1.485   Mean   :15.56   Mean   :14.8
##  3rd Qu.:21.00   3rd Qu.:2.000   3rd Qu.:21.00   3rd Qu.:21.0
```

```
##  Max.   :32.00   Max.   :2.000   Max.   :99.00   Max.   :32.0
##  MUNICIPIO_RES    TIPO_PACIENTE   FECHA_INGRESO      FECHA_SINTOMAS
##  Min.   :  1.00   Min.   :1.000   Length:100000      Length:100000
##  1st Qu.:  7.00   1st Qu.:1.000   Class :character   Class :character
##  Median : 17.00   Median :1.000   Mode  :character   Mode  :character
##  Mean   : 33.44   Mean   :1.138
##  3rd Qu.: 39.00   3rd Qu.:1.000
##  Max.   :999.00   Max.   :2.000
##   FECHA_DEF            INTUBADO     NEUMONIA           EDAD
##  Length:100000      Min.   : 1   Min.   : 1.000   Min.   :  0.00
##  Class :character   1st Qu.:97   1st Qu.: 2.000   1st Qu.: 29.00
##  Mode  :character   Median :97   Median : 2.000   Median : 40.00
##                     Mean   :84   Mean   : 2.426   Mean   : 41.25
##                     3rd Qu.:97   3rd Qu.: 2.000   3rd Qu.: 52.00
##                     Max.   :99   Max.   :99.000   Max.   :120.00
##   NACIONALIDAD      EMBARAZO     HABLA_LENGUA_INDIG    INDIGENA
##  Min.   :1.000   Min.   : 1.00   Min.   : 1.000     Min.   : 1.000
##  1st Qu.:1.000   1st Qu.: 2.00   1st Qu.: 2.000     1st Qu.: 2.000
##  Median :1.000   Median : 2.00   Median : 2.000     Median : 2.000
##  Mean   :1.005   Mean   :48.36   Mean   : 5.992     Mean   : 5.914
##  3rd Qu.:1.000   3rd Qu.:97.00   3rd Qu.: 2.000     3rd Qu.: 2.000
##  Max.   :2.000   Max.   :98.00   Max.   :99.000     Max.   :99.000
##     DIABETES          EPOC             ASMA           INMUSUPR
##  Min.   : 1.00   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.00   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 2.00   Median : 2.000   Median : 2.000   Median : 2.000
##  Mean   : 2.17   Mean   : 2.256   Mean   : 2.238   Mean   : 2.277
##  3rd Qu.: 2.00   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :98.00   Max.   :98.000   Max.   :98.000   Max.   :98.000
##   HIPERTENSION       OTRA_COM      CARDIOVASCULAR      OBESIDAD
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median : 2.000
##  Mean   : 2.118   Mean   : 2.404   Mean   : 2.255   Mean   : 2.121
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :98.000
##  RENAL_CRONICA      TABAQUISMO       OTRO_CASO      TOMA_MUESTRA_LAB
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.:1.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median :1.000
##  Mean   : 2.248   Mean   : 2.187   Mean   : 9.648   Mean   :1.153
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:1.000
##  Max.   :98.000   Max.   :98.000   Max.   :99.000   Max.   :2.000
##  RESULTADO_LAB  TOMA_MUESTRA_ANTIGENO RESULTADO_ANTIGENO CLASIFICACION_FINAL
##  Min.   : 1.0   Min.   :1.000         Min.   : 1.00      Min.   :1.000
##  1st Qu.: 1.0   1st Qu.:2.000         1st Qu.:97.00      1st Qu.:3.000
##  Median : 2.0   Median :2.000         Median :97.00      Median :6.000
##  Mean   :16.3   Mean   :1.913         Mean   :88.75      Mean   :5.263
##  3rd Qu.: 2.0   3rd Qu.:2.000         3rd Qu.:97.00      3rd Qu.:7.000
##  Max.   :97.0   Max.   :2.000         Max.   :97.00      Max.   :7.000
##     MIGRANTE    PAIS_NACIONALIDAD  PAIS_ORIGEN            UCI
##  Min.   : 1.0   Length:100000      Length:100000      Min.   : 1.00
##  1st Qu.:99.0   Class :character   Class :character   1st Qu.:97.00
##  Median :99.0   Mode  :character   Mode  :character   Median :97.00
```

```
##  Mean   :98.6                                          Mean   :84.02
##  3rd Qu.:99.0                                          3rd Qu.:97.00
##  Max.   :99.0                                          Max.   :99.00
```

```
glimpse(covid)
```

```
## Rows: 100,000
## Columns: 40
## $ FECHA_ACTUALIZACION   <chr> "25/12/2020", "25/12/2020", "25/12/2020", "25/12~
## $ ID_REGISTRO           <chr> "3f4171", "2fd222", "83800", "2bc493", "31ed7f",~
## $ ORIGEN                <int> 2, 2, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, ~
## $ SECTOR                <int> 12, 12, 5, 4, 4, 9, 12, 12, 4, 12, 12, 12, 12, 1~
## $ ENTIDAD_UM            <int> 9, 9, 20, 6, 23, 19, 19, 21, 1, 19, 3, 10, 9, 9,~
## $ SEXO                  <int> 2, 1, 2, 2, 2, 2, 1, 1, 2, 1, 2, 1, 2, 1, 2, 2, ~
## $ ENTIDAD_NAC           <int> 9, 9, 15, 14, 30, 19, 19, 9, 1, 19, 24, 9, 9, 9,~
## $ ENTIDAD_RES           <int> 9, 9, 20, 6, 23, 19, 19, 21, 1, 19, 3, 15, 9, 9,~
## $ MUNICIPIO_RES         <int> 5, 7, 160, 10, 5, 39, 6, 114, 1, 39, 8, 57, 6, 2~
## $ TIPO_PACIENTE         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, ~
## $ FECHA_INGRESO         <chr> "20/10/2020", "04/12/2020", "17/04/2020", "24/11~
## $ FECHA_SINTOMAS        <chr> "20/10/2020", "01/12/2020", "14/04/2020", "21/11~
## $ FECHA_DEF             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ INTUBADO              <int> 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, ~
## $ NEUMONIA              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, ~
## $ EDAD                  <int> 38, 78, 27, 42, 44, 61, 39, 41, 29, 32, 60, 46, ~
## $ NACIONALIDAD          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ EMBARAZO              <int> 97, 2, 97, 97, 97, 97, 2, 2, 97, 2, 97, 2, 97, 2~
## $ HABLA_LENGUA_INDIG    <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ INDIGENA              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ DIABETES              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, ~
## $ EPOC                  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ ASMA                  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ INMUSUPR              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ HIPERTENSION          <int> 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, ~
## $ OTRA_COM              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ CARDIOVASCULAR        <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ OBESIDAD              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, ~
## $ RENAL_CRONICA         <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ TABAQUISMO            <int> 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 1, ~
## $ OTRO_CASO             <int> 2, 2, 99, 2, 2, 2, 2, 1, 2, 1, 2, 1, 1, 1, 99, 2~
## $ TOMA_MUESTRA_LAB      <int> 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ RESULTADO_LAB         <int> 2, 97, 2, 2, 1, 2, 2, 2, 1, 1, 2, 2, 3, 1, 1, 2,~
## $ TOMA_MUESTRA_ANTIGENO <int> 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, ~
## $ RESULTADO_ANTIGENO    <int> 97, 1, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 2~
## $ CLASIFICACION_FINAL   <int> 7, 3, 7, 7, 3, 7, 7, 7, 3, 3, 7, 7, 6, 3, 3, 7, ~
## $ MIGRANTE              <int> 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, ~
## $ PAIS_NACIONALIDAD     <chr> "Mexico", "Mexico", "Mexico", "Mexico", "Mexico"~
## $ PAIS_ORIGEN           <chr> "97", "97", "97", "97", "97", "97", "97", "97", ~
## $ UCI                   <int> 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, ~
```

4) Filtra los renglones que dieron positivo para SARS-COVID y calcula el numero de registros Los casos positivos son aquellos que en la columna CLASIFICACION_FINAL tienen 1, 2 o 3

```
covid %>% filter(CLASIFICACION_FINAL < 4) %>% count()
```

```
##       n
## 1 39220
```

```
## Observación. 2/5 de la muestra dieron positivo a covid
```

5) Cuenta el numero de registros nulos por columna (HINT: Usar sapply o map, e is.na)

```
covid %>% sapply(function(y) sum(length(which(is.na(y)))))
```

```
##   FECHA_ACTUALIZACION           ID_REGISTRO                ORIGEN
##                     0                     0                     0
##                SECTOR           ENTIDAD_UM                  SEXO
##                     0                     0                     0
##           ENTIDAD_NAC           ENTIDAD_RES         MUNICIPIO_RES
##                     0                     0                     0
##         TIPO_PACIENTE         FECHA_INGRESO        FECHA_SINTOMAS
##                     0                     0                     0
##             FECHA_DEF              INTUBADO              NEUMONIA
##                 95044                     0                     0
##                  EDAD          NACIONALIDAD              EMBARAZO
##                     0                     0                     0
##    HABLA_LENGUA_INDIG              INDIGENA              DIABETES
##                     0                     0                     0
##                  EPOC                  ASMA              INMUSUPR
##                     0                     0                     0
##          HIPERTENSION              OTRA_COM        CARDIOVASCULAR
##                     0                     0                     0
##              OBESIDAD         RENAL_CRONICA             TABAQUISMO
##                     0                     0                     0
##             OTRO_CASO      TOMA_MUESTRA_LAB          RESULTADO_LAB
##                     0                     0                     0
## TOMA_MUESTRA_ANTIGENO    RESULTADO_ANTIGENO    CLASIFICACION_FINAL
##                     0                     0                     0
##              MIGRANTE     PAIS_NACIONALIDAD            PAIS_ORIGEN
##                     0                     0                     0
##                   UCI
##                     0
```

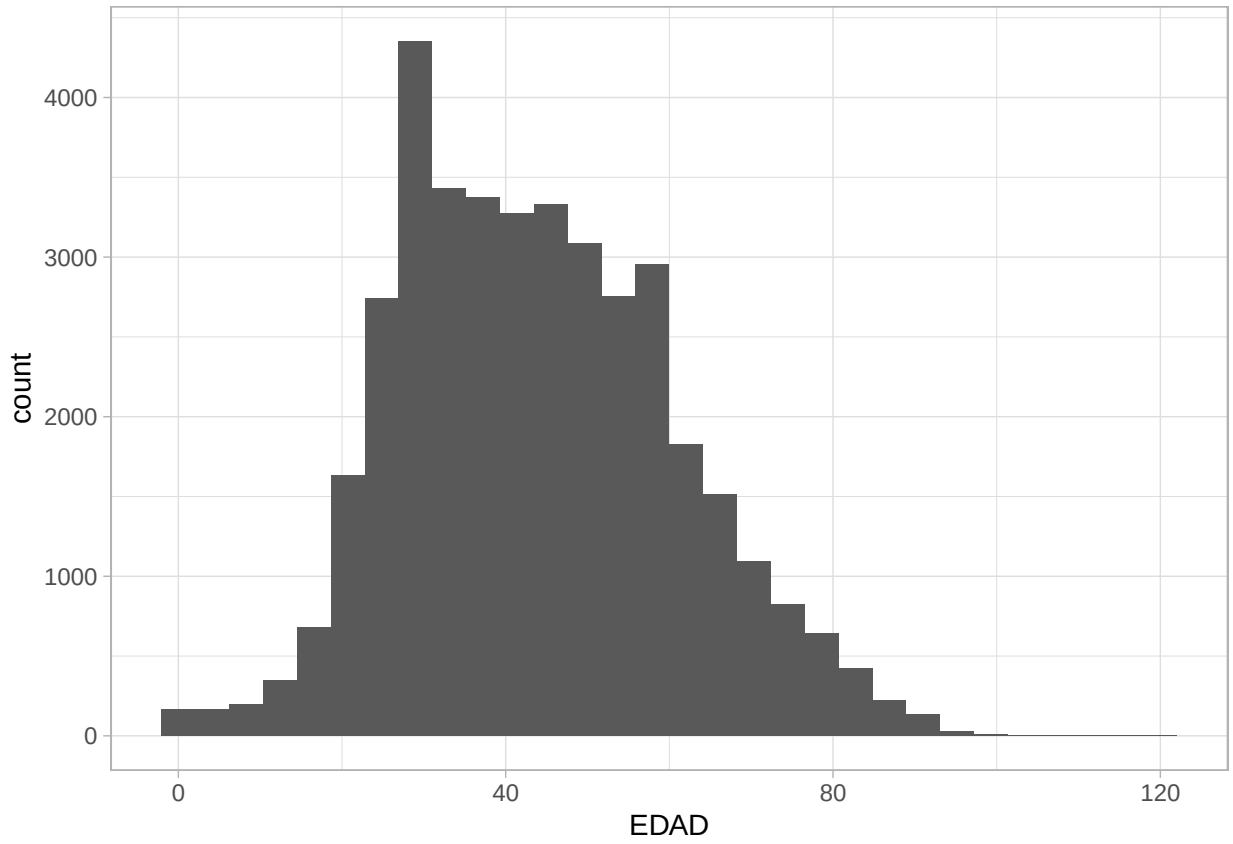
6) a) Calcular la media de edades de los contagiados de covid
   b) Realiza un Histograma de las edades de los contagiados
   c) Realiza una grafica de densidad de edades de los contagiados

```
#a
contagiados <- covid %>% filter(CLASIFICACION_FINAL < 4)
median(contagiados$EDAD)
```

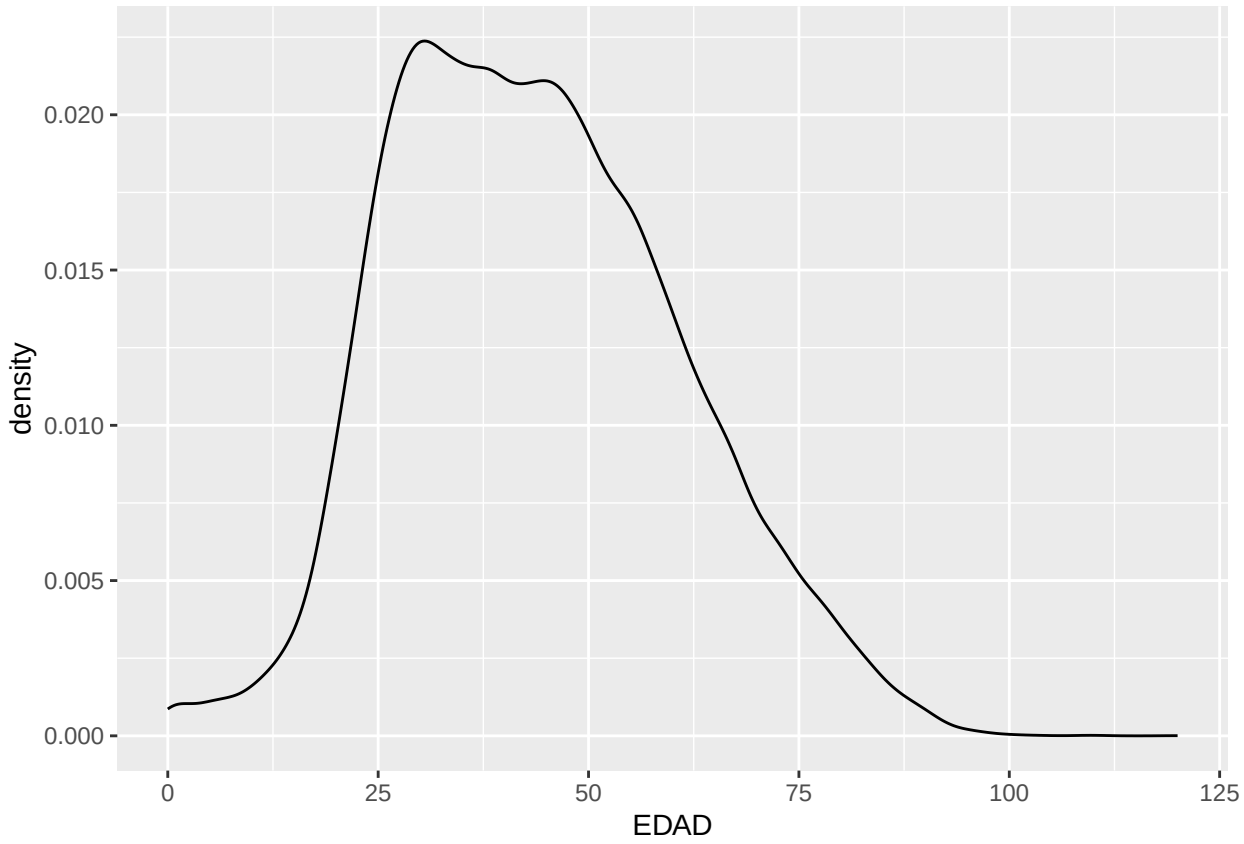
```
## [1] 43
```

```
#b
ggplot(contagiados, aes(x = EDAD)) +
  geom_histogram() + theme_light()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#c
ggplot(contagiados, aes(x = EDAD)) +
  geom_density()
```



7) Agregar una columna nueva al dataframe que tenga valor 1 cuando la fecha de defuncion no es valor nulo y 0 cuando es nulo La columna que contiene la fecha de defuncion se llama FECHA_DEF HINT: Usa mutate, ifelse e is.na

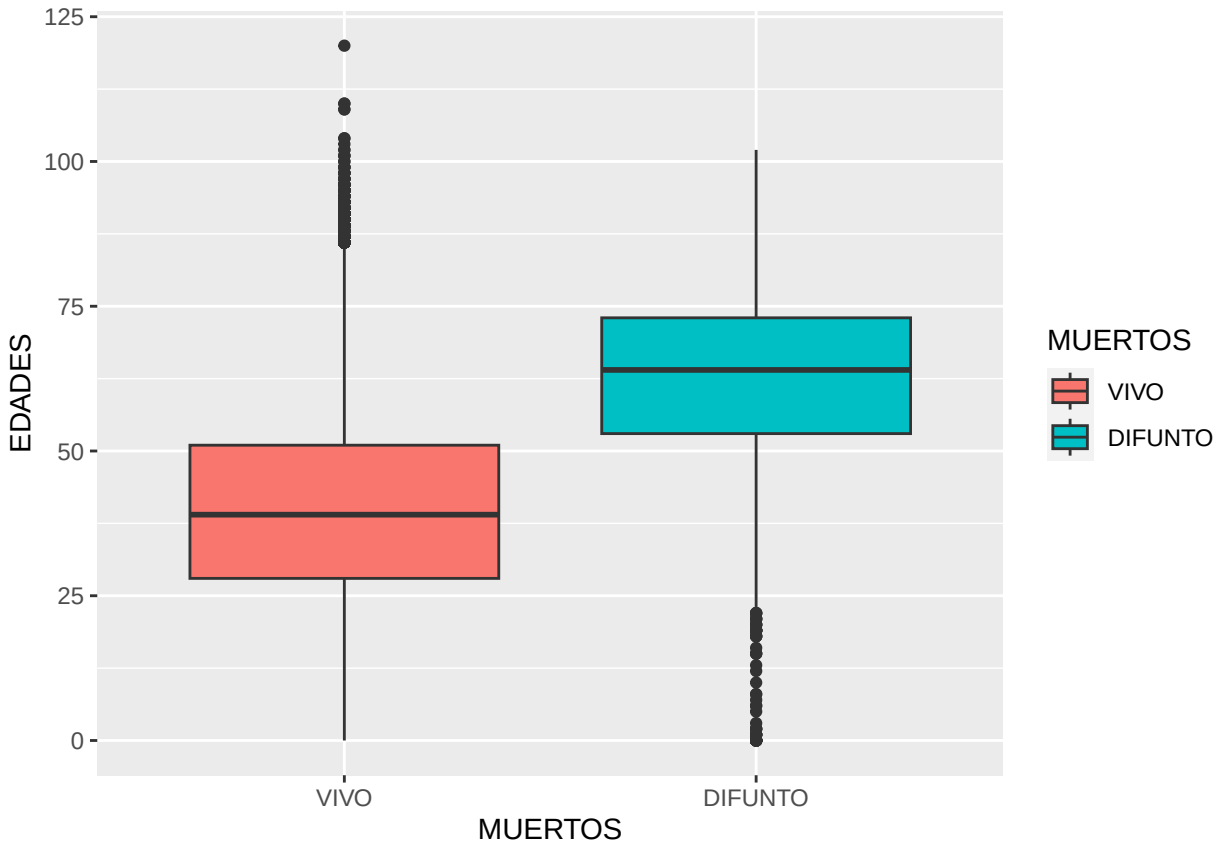
```
covid <- covid %>% mutate(MUERTO = case_when(
            is.na(FECHA_DEF) ~ 0,
            TRUE ~ 1
          ))
```

8) Hacer un boxplot de edades de los muertos por covid vs los que no murieron para ver si detectamos diferencias y escribe tus conclusiones

```
EDADES <- covid$EDAD
MUERTOS <- factor(covid$MUERTO, levels = c(0,1), labels = c("VIVO","DIFUNTO"))
edades <- data.frame(MUERTOS,EDADES)

qplot(x = MUERTOS, y = EDADES, data = edades, geom = "boxplot", fill = MUERTOS)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
edades %>% group_by(MUERTOS) %>% summarize(quants = quantile(EDADES, probs = c(0.25, 0.75)))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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

```
## # A tibble: 4 x 2
## # Groups:   MUERTOS [2]
##   MUERTOS quants
##   <fct>    <dbl>
## 1 VIVO        28
## 2 VIVO        51
## 3 DIFUNTO     53
## 4 DIFUNTO     73
```

##Observación. Es evidente que sí hay una diferencia en la comparación entre aquellos que sobrevivieron a la enfermedad y aquellos que no, ya que aquellos que sobrevivieron al virus en el 50% de los casos son personas menores de 51 años y que las personas que murieron a causa de la enfermedad son mayores a 53 años. ##Por lo tanto, vemos una clara línea divisoria marcada por la edad para aquellas personas que sobreviven de las que no.

9) Transforma la columna CLASIFICACION_FINAL, que tenga valor de 1 si tiene 1, 2 o 3 como valor y

que tenga 0 en cualquier otro caso HINT: Usar transform o mutate

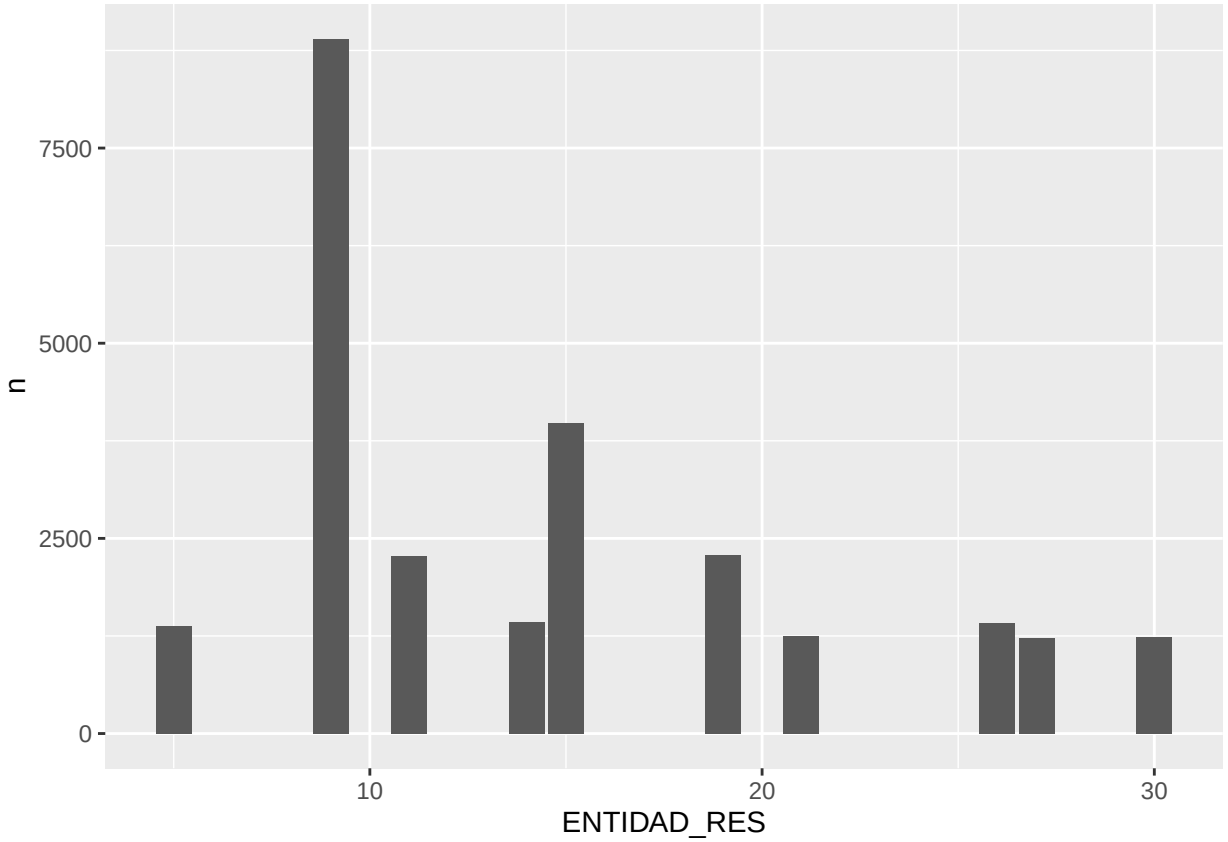
```
covid <- covid %>% transform(CLASIFICACION_FINAL = ifelse(
  CLASIFICACION_FINAL == 1 | CLASIFICACION_FINAL == 2 | CLASIFICACION_FINAL == 3, 1, 0))
```

10) Cuenta el numero de casos positivos agrupado por estado y realiza una grafica de barras de los 10 estados con mas casos # HINT: Usar groupby, summarize, n(), y ggplot2

```
estados_contagiados <- contagiados %>% group_by(ENTIDAD_RES) %>% tally() %>% arrange(desc(n)) %>% top_n
```

```
## Selecting by n
```

```
ggplot(estados_contagiados, aes(x = ENTIDAD_RES, y = n)) +
  geom_bar(stat="identity")
```



11) Renombra la columna llamada CLASIFICACION FINAL para que ahora su nombre sea: "CONTAGIADO"

```
covid <- covid %>% rename(CONTAGIADO = CLASIFICACION_FINAL)
```

12) Realiza una funcion que al aplicarla nos diga el procentaje del total de registros que estan contagiados por Covid Ejemplo: al correr la funcion porcentaje_contagios(mi_dataframe) el resultado sea: 20.5%

```
porcentaje_contagiados <- function(mi_dataframe, columna) {
  total <- mi_dataframe %>% tally()
  contagios <- mi_dataframe %>% filter(columna == 1) %>% tally()
  ifelse(contagios == 0, paste(0, "%"), paste((contagios/total)*100, "%"))
}
porcentaje_contagiados(covid, covid$CONTAGIADO)
```

```
##      n
```

```
## [1,] "39.22 %"
```

13) Realiza una matriz de corrrelacion entre las variables numericas y concluye HINT: https://stackoverflow.com/questions/5863097/selecting-only-numeric-columns-from-a-data-frame

```
var_num <- select_if(covid, is.numeric)

# Correlacionados
correlacion <- cor(var_num)
j = 1
for(i in correlacion){
  if(i > -0.5 & i < 0.5){
    correlacion[j] = ""
  }
  j = j + 1
}

# Poco Correlacionados
poca_correlacion <- cor(var_num)
k = 1
for(i in poca_correlacion){
  if((i > -0.5 & i < -0.25) | (i > 0.25 & i < 0.5)){
    poca_correlacion[k] = poca_correlacion[k]
  }else{
    poca_correlacion[k] = 0
  }
  k = k + 1
}


#NOTAS: una correlación positiva perfecta: r = 1. (cuando una sube la otra sube) una correlación negati
#Observación:
  #Las variables más correlacionadas son:
   #1.- Las complicaciones presentadas como:
     #diabetes - EPOC - asma - inmunosupresores - hipertensión -
     #problemas cardiovasculares - obesidad - problemas renales - tabaquismo
   #2.- La correlación entre pacientes intubados y el número de muertes
#Las variables poco correlacionadas son: La edad correlacionada con intubación y los decesos
```

CONCLUSIONES: Pese a que actualmente mucha de esta información no queda clara. Si está fuera la primera vez que se sacan conclusiones con respecto al estudio de estos datos estos serían mis comentarios.

Un factor importante hablando de probabilidades de supervivencia es mas alta para aquellos menores de 50 a?os. Tener complicaciones como la diabetes, la hipertención, asma, problemas cardiovasculares,problemas de obesidad, el uso de inmunosupresores,tener precedentes de tabaquismo, problemas renales crónicos, y algunas otras complicaciones aumenta el riesgo de la enfermedad. Por otro lado, existe tambien una correlación entre las personas intubadas y los decesos.