R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```
library(readr)
library(dplyr)
library(lubridate) #fechas
```

```
D08 <-read_csv("csv_files/2021-08.csv",col_types = cols())
D09 <-read_csv("csv_files/2021-09.csv",col_types = cols())
D10 <- read_csv("csv_files/2021-10.csv",col_types = cols())
dataset<- do.call("rbind", list(D08,D09,D10))
```

```
head(dataset)
```

```
## # A tibble: 6 x 9
##   Genero_Usuario Edad_Usuario  Bici Ciclo_Estacion_Ret~ Fecha_Retiro Hora_Retiro
##   <chr>                 <dbl> <dbl>               <dbl> <chr>        <time>
## 1 M                        28 11302                  73 31/07/21     23:57:44
## 2 M                        33 10571                 121 31/07/21     23:54:00
## 3 M                        19 12451                 132 31/07/21     23:52:58
## 4 M                        32  8314                   7 31/07/21     23:23:41
## 5 F                        36  7993                   7 31/07/21     23:23:17
## 6 M                        34 10675                 136 31/07/21     23:56:42
## # ... with 3 more variables: Ciclo_EstacionArribo <dbl>, 'Fecha Arribo' <chr>,
## #   Hora_Arribo <time>
```

```
summary(dataset)
```

```
##  Genero_Usuario      Edad_Usuario        Bici       Ciclo_Estacion_Retiro
##  Length:1103273     Min.   :17.00   Min.   :  775   Min.   :   1.0
##  Class :character   1st Qu.:28.00   1st Qu.: 7892   1st Qu.:  70.0
##  Mode  :character   Median :33.00   Median : 9456   Median : 159.0
##                     Mean   :35.96   Mean   : 9425   Mean   : 187.3
##                     3rd Qu.:41.00   3rd Qu.:11212   3rd Qu.: 290.0
##                     Max.   :86.00   Max.   :15339   Max.   :3002.0
##  Fecha_Retiro       Hora_Retiro       Ciclo_EstacionArribo Fecha Arribo
##  Length:1103273     Length:1103273    Min.   :  1.0        Length:1103273
##  Class :character   Class1:hms        1st Qu.: 68.0        Class :character
##  Mode  :character   Class2:difftime   Median :154.0        Mode  :character
##                     Mode  :numeric    Mean   :184.5
##                                       3rd Qu.:285.0
```

```
##                                       Max.   :480.0
##  Hora_Arribo
##  Length:1103273
##  Class1:hms
##  Class2:difftime
##  Mode  :numeric
##
##
```

Conclusiones: Necesitamos cambiar los tipos de fecha de retiro/arribo a Date En la pág de ecobici, podemos ver que las cicloestaciones activas son 480, por lo que en estación retiro tenemos outliers.

Errores en Fechas

```
#Formatear fecha
fecha_r<-mdy(dataset$Fecha_Retiro)
```

```
## Warning: 694129 failed to parse.
```

! Tenemos diferencias de escritura en algunas fechas

```
#Verificar número de digitos en ambas fechas
dataset %>% count(nchar(dataset$Fecha_Retiro))
```

```
## # A tibble: 2 x 2
##   `nchar(dataset$Fecha_Retiro)`      n
##                           <int>  <int>
## 1                             8 343183
## 2                            10 760090
```

```
dataset %>% count(nchar(dataset$`Fecha Arribo`))
```

```
## # A tibble: 2 x 2
##   `nchar(dataset$\`Fecha Arribo\`)`      n
##                               <int>  <int>
## 1                                 8 343183
## 2                                10 760090
```

```
head(dataset[nchar(dataset$Fecha_Retiro)==8,])
```

```
## # A tibble: 6 x 9
##   Genero_Usuario Edad_Usuario  Bici Ciclo_Estacion_Ret~ Fecha_Retiro Hora_Retiro
##   <chr>                 <dbl> <dbl>               <dbl> <chr>        <time>
## 1 M                        28 11302                  73 31/07/21     23:57:44
## 2 M                        33 10571                 121 31/07/21     23:54:00
## 3 M                        19 12451                 132 31/07/21     23:52:58
## 4 M                        32  8314                   7 31/07/21     23:23:41
## 5 F                        36  7993                   7 31/07/21     23:23:17
## 6 M                        34 10675                 136 31/07/21     23:56:42
## # ... with 3 more variables: Ciclo_EstacionArribo <dbl>, `Fecha Arribo` <chr>,
## #   Hora_Arribo <time>
```

```
head(dataset[nchar(dataset$Fecha_Retiro)==10,])
```

```
## # A tibble: 6 x 9
##   Genero_Usuario Edad_Usuario  Bici Ciclo_Estacion_Ret~ Fecha_Retiro Hora_Retiro
##   <chr>                 <dbl> <dbl>               <dbl> <chr>        <time>
## 1 M                        63  8642                  47 31/08/2021   23:51:30
## 2 M                        45  9389                 138 01/09/2021   00:02:43
## 3 F                        36 10279                  65 31/08/2021   23:27:14
## 4 M                        49 12418                   8 31/08/2021   23:18:30
## 5 M                        24  8732                  10 31/08/2021   23:51:29
## 6 M                        34 12439                   9 01/09/2021   00:11:07
## # ... with 3 more variables: Ciclo_EstacionArribo <dbl>, `Fecha Arribo` <chr>,
## #   Hora_Arribo <time>
```

- La escritura de año es lo que esta cambiando, dado que todos los años deben ser 2021, es facil cambiar todos.

```
#Cambiar por 2021 todos los años
#Fecha Retiro:
dataset$Fecha_Retiro<- paste(substr(dataset$Fecha_Retiro,0,6),"2021",sep="")

#Fecha Arribo:
dataset$`Fecha Arribo`<- paste(substr(dataset$`Fecha Arribo`,0,6),"2021",sep="")
```

```
#Formatear a Fecha
dataset$Fecha_Retiro<-as.Date(dataset$Fecha_Retiro,format="%d/%m/%Y")
dataset$`Fecha Arribo`<-as.Date(dataset$`Fecha Arribo`,format="%d/%m/%Y")
```

Outliers estación Arribo

```
#Ver cuantas estaciones no están en los rangos.
dataset[dataset$Ciclo_Estacion_Retiro>480,]
```

```
## # A tibble: 2 x 9
##   Genero_Usuario Edad_Usuario  Bici Ciclo_Estacion_Ret~ Fecha_Retiro Hora_Retiro
##   <chr>                 <dbl> <dbl>               <dbl> <date>       <time>
## 1 M                        26  9592                3002 2021-08-12   08:41:08
## 2 M                        26 11740                3002 2021-10-14   08:29:15
## # ... with 3 more variables: Ciclo_EstacionArribo <dbl>, `Fecha Arribo` <date>,
## #   Hora_Arribo <time>
```

Tenemos solo 2 outliers, dado que parece ser un error manual, optare por revisar en que estacion tenemos más afluencia

```
dplyr::count(dataset, Ciclo_Estacion_Retiro, sort = TRUE) %>% mutate(percentage = n/sum(n))%>% filter(C
```

```
## # A tibble: 2 x 3
##   Ciclo_Estacion_Retiro     n percentage
##                   <dbl> <int>      <dbl>
## 1                   302  3389    0.00307
## 2                   300  1783    0.00162
```

Optare por añadir a los usuarios a la estación 302

```
dataset$Ciclo_Estacion_Retiro[which(dataset$Ciclo_Estacion_Retiro == 3002)] <- 302
```

```
#Verificar datos
dplyr::count(dataset, Ciclo_Estacion_Retiro, sort = TRUE) %>% mutate(percentage = n/sum(n))%>% filter(C
```

```
## # A tibble: 2 x 3
##   Ciclo_Estacion_Retiro     n percentage
##                   <dbl> <int>      <dbl>
## 1                   302  3391    0.00307
## 2                   300  1783    0.00162
```

Guardar Archivo

```
write.csv(dataset, file = "csv_files/dataset.csv")
```