**Estadística y Programación con R 2020**

**\_\_\_**

**“Breve análisis estadístico e interpretación de un conjunto de datos (Covid - 19)”**

**Checkpoint**

**Alumno:** Céspedes Jaén Abdiel

**Experto:** Franco Jenner

**Grupo:** data - analysis - gdl - 20 - 06

**Fecha de entrega:** Sábado 05/12/2020

**ÍNDICE**

[**Resumen**](#_obnk8wqj7ikm) **4**

[**Objetivos**](#_2qymby987lz8) **4**

[Objetivo general](#_kh2427q32atz) 4

[Objetivos particulares](#_hrcuy0w9m9yj) 4

[**Contexto**](#_h9p7jixdrbg3) **5**

[Tipo de datos](#_a73uz7juzsa8) 5

[**Desarrollo**](#_y7hfxi372t21) **5**

[Hipótesis](#_rhnc4dbcyphx) 5

[Preguntas principales](#_rz1ueallbn66) 6

[Procedimiento](#_wre88rkuec7f) 6

[Post Work 2](#_4m58rxy4pim0) 6

[*Importar el dataset a RStudio (Uso de read.csv())*](#_sugp42cizztu) *6*

[*Instalar paquetería*](#_1fjct6tzhlhy) *6*

[Post Work 5](#_oqrfu2iy58lz) 6

[*Determinar la estructura del dataframe (Uso de str())*](#_q81y2fg5wnm1) *6*

[*Cambiar nombres de las columnas*](#_wtm5r2x6bx8m) *6*

[*Selección de valores únicos*](#_ibr286zcxynd) *7*

[*Eliminación de columnas innecesarias*](#_fkom7cf1ai3d) *7*

[*Mínimo, máximo, media y mediana*](#_3y591inlc91e) *7*

[Proyecto Final A](#_ijnc33x5pq9u) 7

[*Repositorio*](#_wrzfi9863qby) *7*

[*Fecha Seleccionada*](#_nvh7eynsucdj) *8*

[*Modo raw*](#_fzb3jn1y3sot) *9*

[*Uso de mean() para las columnas numéricas*](#_scaqmmkb5djf) *9*

[*Uso de summary()*](#_wugs9nm54yh0) *9*

[*Tabla de frecuencia para la columna de “país”*](#_pamp1hyf1w0k) *10*

[*Columna DeathRate*](#_lnppk8wrf8uk) *10*

[*Columna RecoveryRate*](#_kr2hco8wq4tm) *10*

[*Matriz de correlación*](#_39sagmmq746) *10*

[*Gráfica de tabla de frecuencia*](#_1rhbnze22d1g) *11*

[*Filtrado de DeathRate y de RecoveryRate (de mayor a menor y solo 20)*](#_z05zgn1vgcdy) *12*

[*Histogramas de las columnas filtradas*](#_73lm9od3j511) *12*

[Proyecto Final B](#_nzains4r35v3) 13

[*Muestra aleatoria simple del 5% de la población*](#_x4yv5275okxx) *13*

[*Nuevo dataframe*](#_1pqw49o4ivqn) *13*

[*Visualización de la distribución de DeathRate y RecoveryRate de la MUESTRA*](#_ymkauq1g9enq) *13*

[*Visualización de la distribución de DeathRate y RecoveryRate de la POBLACIÓN*](#_u4kmwxblelh1) *14*

[*Comparación de visualizaciones*](#_5qtbn6z64381) *14*

[Proyecto Final C](#_xta7bxggflny) 14

[*Lectura de archivos 18 - 23 de marzo 2020*](#_c8alrsfn5ehq) *14*

[*Dataframe único (Uso de rbind())*](#_y9koamvjsnuw) *15*

[*Filtrado para Estados Unidos*](#_3ohuabrbowth) *16*

[*Visualización a lo largo del tiempo*](#_196ktghspo0h) *16*

[*Exponencial o no*](#_dagqi9zahba) *16*

[Post Work 6](#_n0f5ser3l3yj) 17

[*Estructura del dataframe de Estados Unidos*](#_pq8uk1nmghx0) *17*

[*Revisión de NaN*](#_c8s9rxcmb60l) *18*

[*Variables numéricas elegidas*](#_wptb7ozbrgm5) *18*

[*Histogramas con medias y medianas*](#_mgxis6kixi29) *18*

[*Muestra aleatoria sin reemplazo del 1%*](#_b4l6kbhqytt8) *19*

[*Histogramas de la muestra*](#_uhxckd8vz5vs) *19*

[*Búsqueda de distribución normal*](#_z8m1qvmfvpom) *20*

[Regresión lineal](#_ajq8zf55l5im) 21

[Respuestas](#_uqwln5iya21a) 22

[**Conclusión**](#_vqf03f6217e8) **22**

[**Rúbrica**](#_ix6lw0g8ss42) **23**

[Elementos contemplados en la realización de este proyecto](#_l4rhvuqh0tp9) 23

[**Bibliografía**](#_iax17xnhuqmh) **24**

# 

# Resumen

En el presente trabajo se desglosa el procedimiento empleado por el alumno para dar respuesta a un conjunto de interrogantes realizadas con respecto a un conjunto de datos. La principal intención es emplear las diversas herramientas que ofrece el lenguaje de programación de R para que, sumado además con el uso del programa de R Studio, se pueda dar una solución justificada a las interrogantes planteadas con una justificación desde un punto de vista estadístico.

El proyecto realizado fue el proporcionado en el repositorio de trabajo de la plataforma de BEDU, desglosado, a su vez, en tres partes principales y complementadas por un conjunto de Post Works. No todos los Post Works estaban enfocados a la solución del proyecto final del repositorio, por lo que el procedimiento mostrado a lo largo del trabajo corresponde - además de a “Proyecto Final A”, “B” y “C” - a Post Work 2, Post Work 5, Post Work 6 y Post Work 7 (este último, al estar más enfocado al *storytelling,* entra de lleno en la presentación complementaria, pero también en base a ese mismo Post Work se basó la estructura lógica de este trabajo).

# Objetivos

## Objetivo general

*“Al finalizar el módulo serás capaz de aplicar conocimientos de estadística aplicada y del lenguaje de programación R con el objetivo de generar análisis estadísticos o predictivos, consumiendo datos de bases de datos en la nube, compartiendo los resultados mediante dashboards en una aplicación Web de autoría propia.”*

## Objetivos particulares

* Poner en práctica y en evidencia los conocimientos adquiridos a lo largo del curso.
* Dar respuesta a las interrogantes señaladas a lo largo del proyecto.
* Justificar las respuestas a cada pregunta del proyecto mediante un análisis estadístico.
* Formar un adecuado *storytelling* para exponer los hallazgos encontrados procurando:
  + Enfocar a la audiencia.
  + Enfatizar el objetivo.
  + Crear una narrativa envolvente.

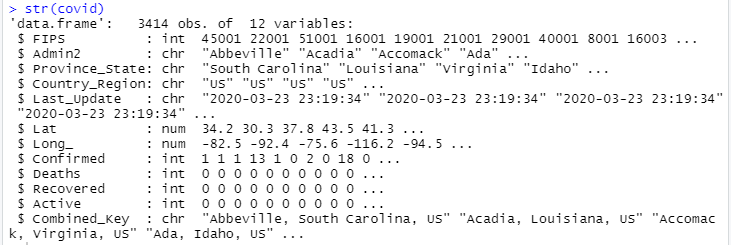
# Contexto

*“A finales del 2019 se desató una pandemia, el COVID-19. Queremos analizar las situaciones de los distintos países en algunos días de interés. En el repositorio* [*https://github.com/CSSEGISandData/COVID-19/tree/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports*](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports)*, podemos encontrar información que nos ayudará a realizar nuestros datos.”*

## Tipo de datos

Los datos de la base de datos del archivo “03-23-2020.csv” se componen de tres tipos de variables: enteros, caracteres y numéricos. Además, gracias a la sentencia *str()* podemos saber, además de los nombres de las columnas que componen el dataframe, que se compone de 3414 observaciones/registros y que hay 12 variables/columnas.

Asimismo, cabe señalar que los datos representan, en orden por el que aparecen en la imagen de abajo, lo siguiente: FIPS (que yo interpreto como una especie de ID), Admin2 (no sé qué interpretación darle), la provincia o el estado, el país o la región, la última fecha de actualización, la latitud, la longitud, los casos confirmados, las muertes confirmadas, los casos recuperados, los casos activos y un dato que menciona como llave. Debo mencionar, como complemento, que no todos los archivos del repositorio contienen la misma estructura, razón por la cual tuve que hacer intervenir para unificar todos los dataframes en uno solo mediante *rbind().*



# Desarrollo

## Hipótesis

*“Los datos de las variables obtenidos de las muestras de los registros de Covid - 19 tienden a distribuirse exponencialmente.”*

## Preguntas principales

1. ¿Es Estados Unidos el país con mayor número de casos confirmados hasta la fecha en que fue tomada la información?
2. ¿La muestra aleatoria simple del 5% de la población sí es representativa de la misma?
3. ¿Cuáles son las variables del dataframe de la muestra de registros de covid que más se correlacionan?
4. ¿La variable de los casos recuperados se distribuye de manera normal?

## Procedimiento

### Post Work 2

#### Importar el dataset a RStudio (Uso de *read.csv()*)



#### Instalar paquetería



### Post Work 5

#### Determinar la estructura del dataframe (Uso de *str()*)

#### 

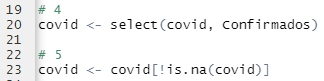
#### Cambiar nombres de las columnas



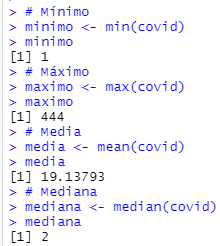
#### Selección de valores únicos

#### 

#### Eliminación de columnas innecesarias



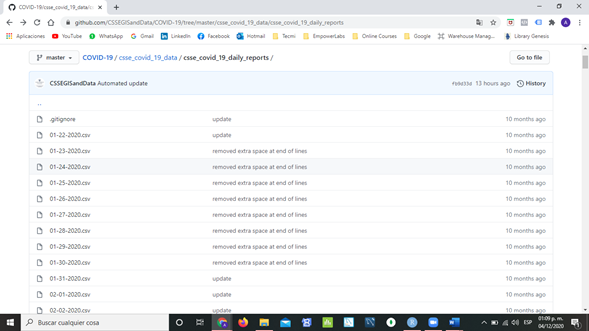
#### Mínimo, máximo, media y mediana



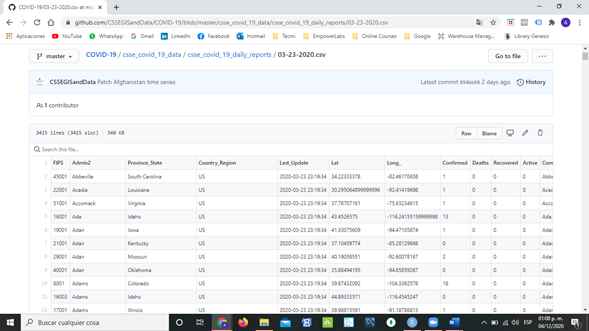
### Proyecto Final A

#### Repositorio

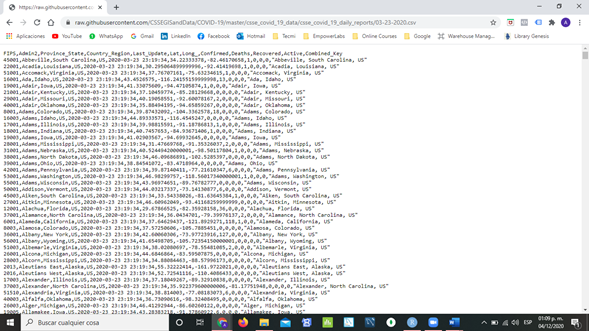
El presente trabajo se realizó en base a la información obtenida del espacio de trabajo cuya liga se presenta a continuación. **Liga:** <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports>



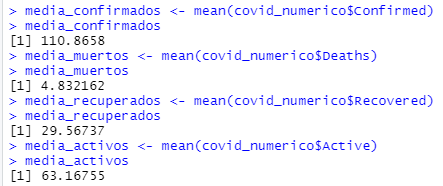
#### Fecha Seleccionada



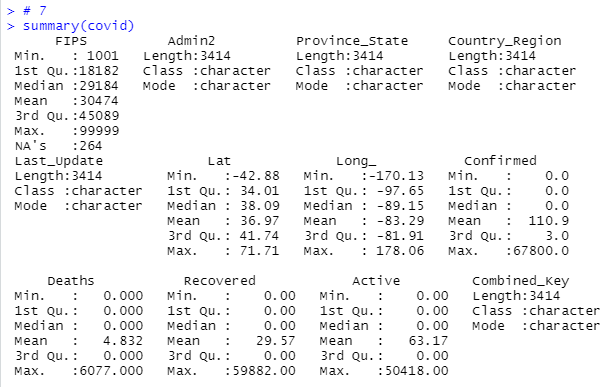
#### Modo raw



#### Uso de *mean()* para las columnas numéricas



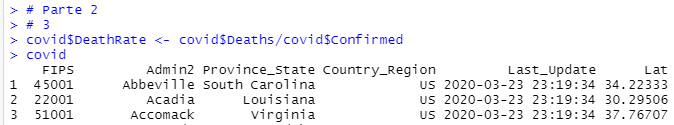
#### Uso de *summary()*



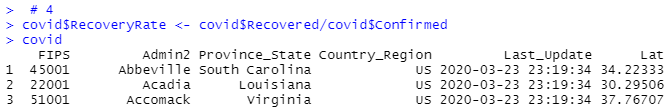
#### Tabla de frecuencia para la columna de “país”



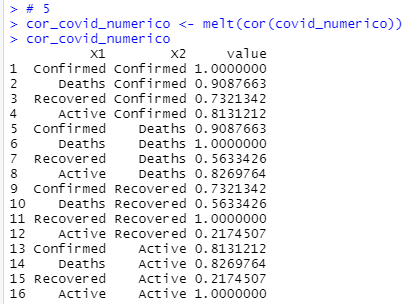
#### Columna *DeathRate*

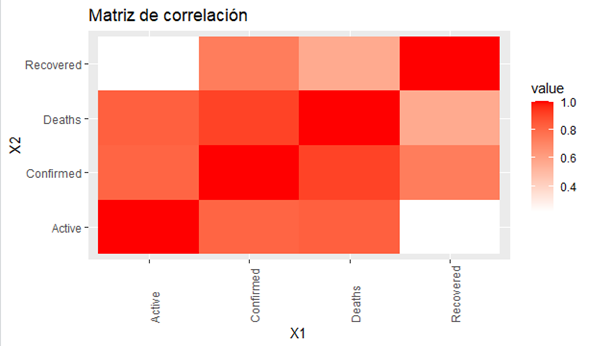


#### Columna *RecoveryRate*



#### Matriz de correlación

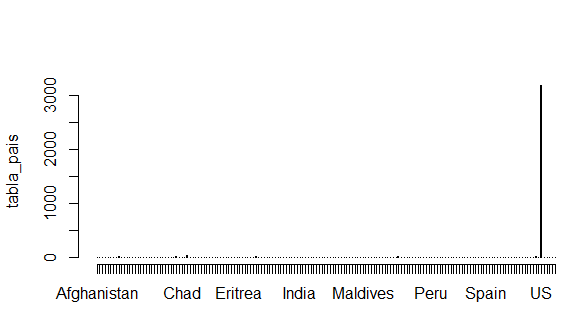




**Descripción:** Gracias al gráfico de la matriz de correlación podemos apreciar que las variables más correlacionadas son los “casos confirmados” con las “muertes”. Numéricamente, vemos que esta misma correlación tiene un valor de 0.908, muy cercano a 1, lo cual nos quiere decir que es una correlación positiva bastante fuerte.

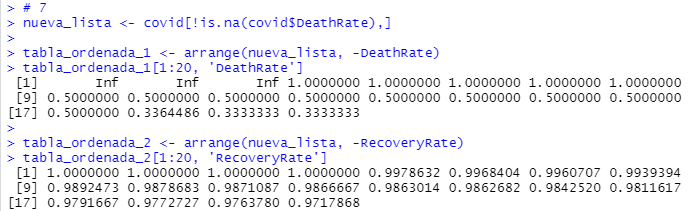
#### Gráfica de tabla de frecuencia



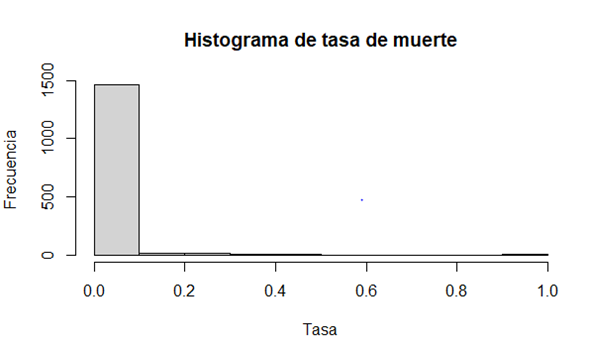


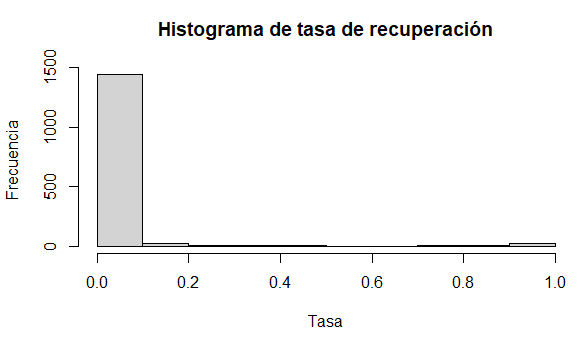
**Descripción:** Gracias a la gráfica de la tabla de frecuencias, vemos que, en efecto, Estados Unidos es el país con más casos confirmados de personas que padecen covid - 19 y este mismo país dista muchísimo de ser alcanzado por el resto de los países analizados.

#### Filtrado de *DeathRate* y de *RecoveryRate* (de mayor a menor y solo 20)



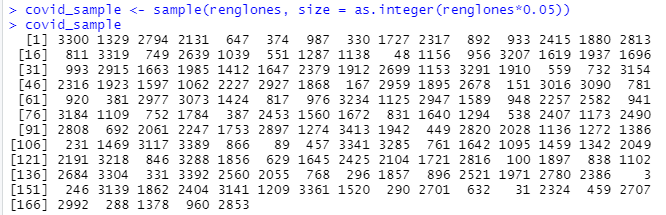
#### Histogramas de las columnas filtradas



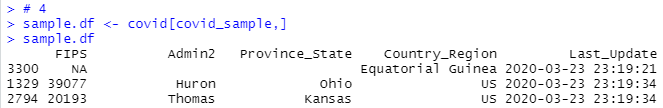


### Proyecto Final B

#### Muestra aleatoria simple del 5% de la población

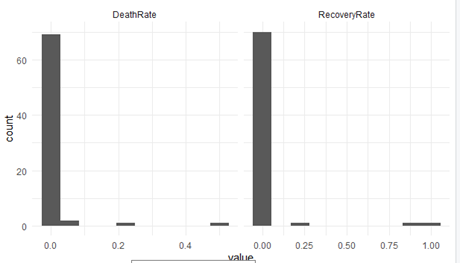


#### Nuevo dataframe



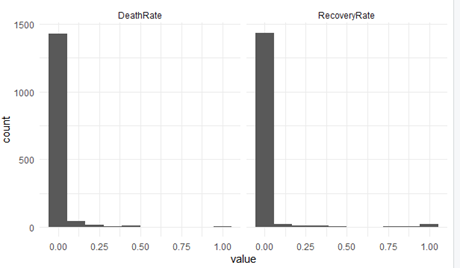
#### Visualización de la distribución de *DeathRate* y *RecoveryRate* de la MUESTRA





#### Visualización de la distribución de *DeathRate* y *RecoveryRate* de la POBLACIÓN





#### Comparación de visualizaciones

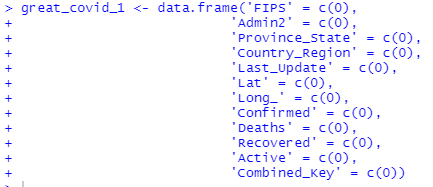
**Descripción:** Tras contrastar visualmente los resultados gráficos de la muestra contra los de la población, he de concluir que la muestra sí es representativa de la población, pues se ve a primera vista hacia dónde tienden a agruparse la mayoría de los datos y que tanto muestra como población tienden a agruparse de igual manera.

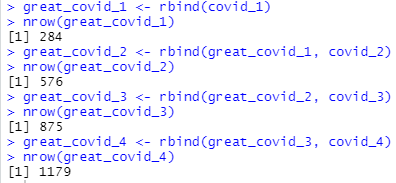
### Proyecto Final C

#### Lectura de archivos 18 - 23 de marzo 2020

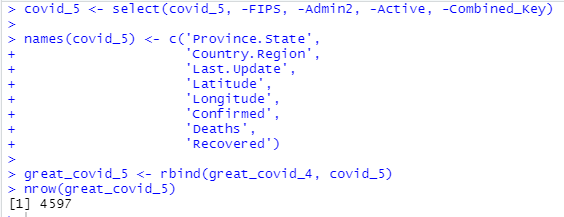


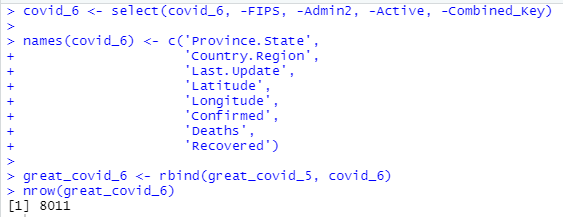
#### Dataframe único (Uso de *rbind()*)





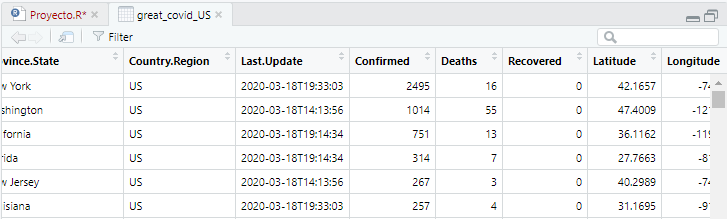
Debido a que ni *covid\_5* ni *covid\_6* estaban estructuradas de las misma manera en que sus predecesoras, tuve que ajustar los nombres para que se pudiera hacer coincidir los datos de unos con los de otros en un único dataframe.



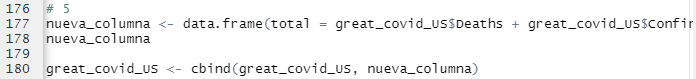


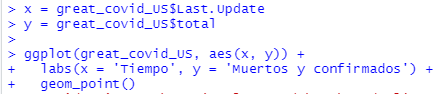
#### Filtrado para Estados Unidos

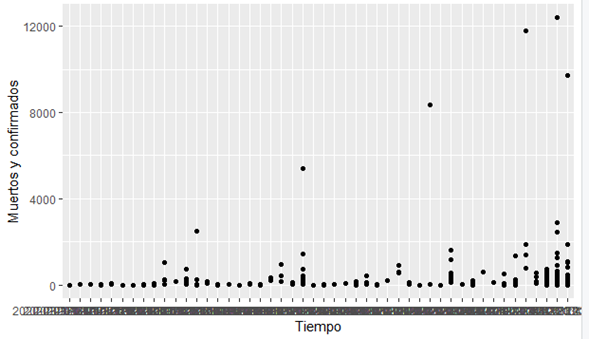




#### Visualización a lo largo del tiempo





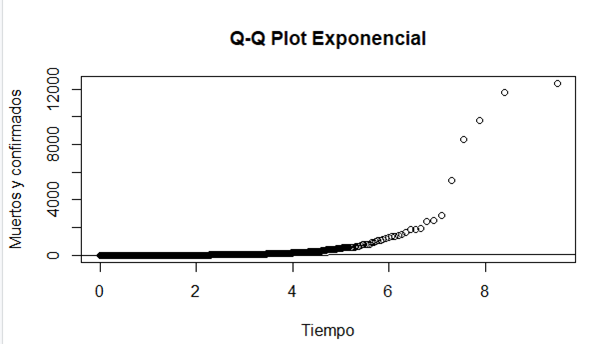


#### Exponencial o no

H0: Los datos del dataframe se distribuyen de manera exponencial.

Ha: Los datos del dataframe NO se distribuyen de manera exponencial.

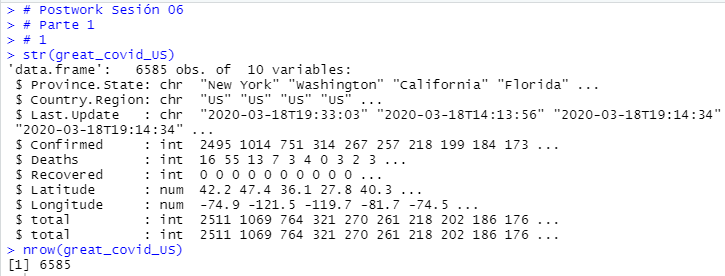




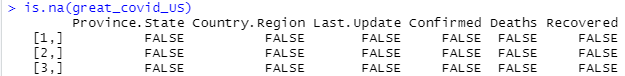
**Descripción:** Grácias a la *quantile - quantile plot,* podemos concluir que, “de momento”, los datos no se distribuyen de manera exponencial, por lo tanto, se rechaza H0 y no se rechaza Ha. Hago énfasis al escribir “de momento” ya que es posible que, debido a la poca cantidad de datos que conforman la muestra que usé para realizar este ejercicio, puede que no tengamos aún suficiente información como para descartar la posibilidad de un comportamiento exponencial.

### Post Work 6

#### Estructura del dataframe de Estados Unidos



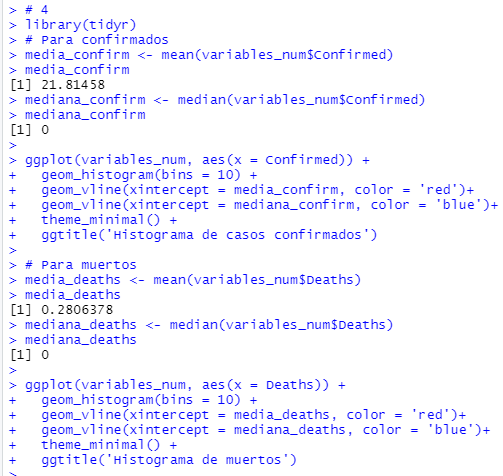
#### Revisión de NaN



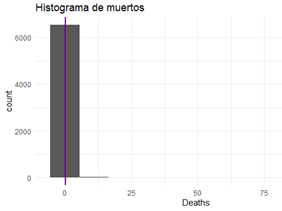
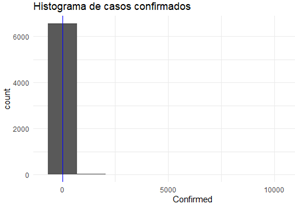
#### Variables numéricas elegidas

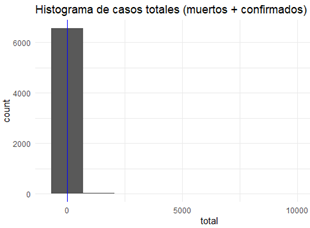


#### Histogramas con medias y medianas

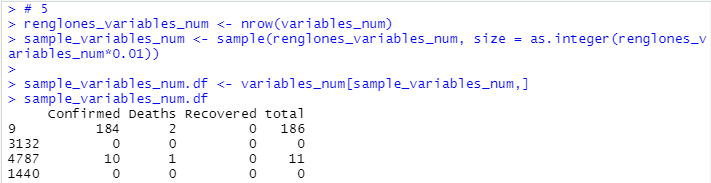




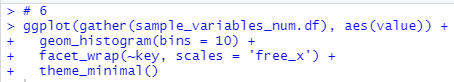


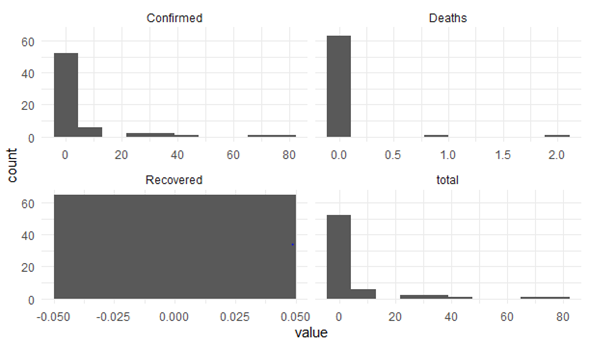


#### Muestra aleatoria sin reemplazo del 1%



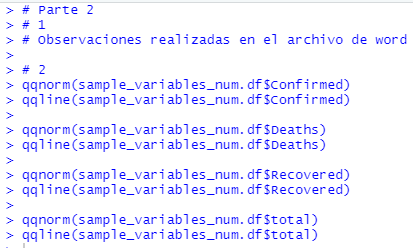
#### Histogramas de la muestra

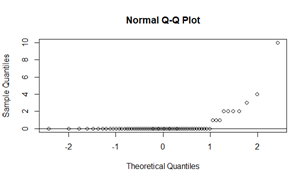
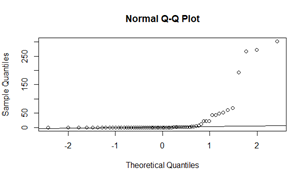


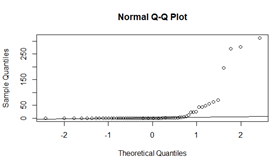
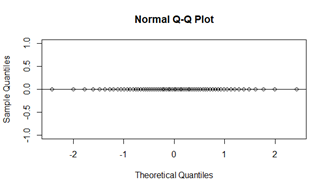


**Descripción:** Previo a la la *qqplot,* supuse que la tercera gráfica (esquina inferior izquierda) se distribuía de manera uniforme, más adelante obtuve otro hallazgo.

#### Búsqueda de distribución normal

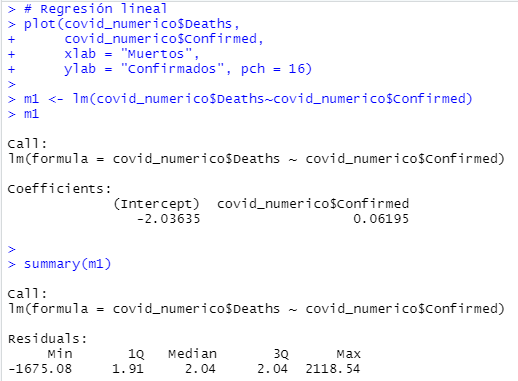


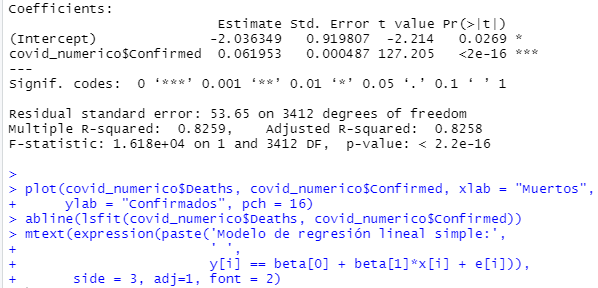


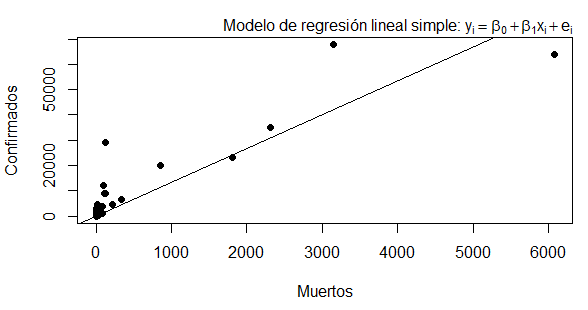


**Descripción:** De todas las gráficas, aquella que creía que se distribuía de manera uniforme también resultó que cumple con los requisitos para considerar que se distribuye de manera normal dado que los puntos - que representan las intersecciones cuartil - cuartil - están totalmente alineados. El resto de variables no presentan una distribución normal.

### Regresión lineal







**Descripción:** Añado un ejemplo de regresión lineal simple para comprobar la relación lineal entre las variables que previamente se vieron más correlacionadas en la matriz homónima. Podemos apreciar gráficamente cómo es la tendencia de ambas variables a corresponderse de manera poco perceptible, pero ya comprobada gracias a la matriz.

## Respuestas

1. ¿Es Estados Unidos el país con mayor número de casos confirmados hasta la fecha en que fue tomada la información?

Sí, Estados Unidos es de lejos el país con mayor número de contagios.

1. ¿La muestra aleatoria simple del 5% de la población sí es representativa de la misma?

Sí, pues presenta un comportamiento de distribución similar al de la población de donde fue extraída.

1. ¿Cuáles son las variables del dataframe de la muestra de registros de covid que más se correlacionan?

La variable de casos confirmados con la variable de muertos.

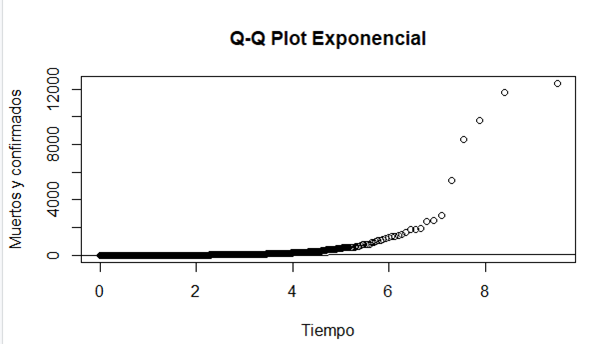
1. ¿La variable de los casos recuperados se distribuye de manera normal?

Sí, gracias a la prueba de *qq plot* podemos comprobar que sí se distribuye de manera normal.

# Conclusión

H0: Los datos del dataframe se distribuyen de manera exponencial.

Ha: Los datos del dataframe NO se distribuyen de manera exponencial.



Nuevamente recurrimos a la ayuda de la *qq plot* para evaluar si la distribución de los datos que corresponden al dataframe de personas que padecen covid - 19 en Estados Unidos se distribuyen exponencialmente y resulta ser que no, dado que los últimos puntos distan bastante de la línea que marca la referencia con respecto a los cuantiles teóricos de una distribución exponencial.

Por lo tanto, se rechaza H0, que fue la hipótesis inicialmente planteada y se no se rechaza Ha; sin embargo, nuevamente hago énfasis que esto es con el análisis de los datos que emplee para hacer esta evaluación.

# Rúbrica

## Elementos contemplados en la realización de este proyecto

|  |  |
| --- | --- |
| **Sesión** | **Elementos** |
| I. Curso propedéutico introducción a R | Declarar y acceder a vectores |
| Declarar y manipular dataframes |
| Leer datos desde diferentes fuentes |
| Usar la librería *“dplyr”* para manipular datos |
| II. Programación: Funciones, Loops y Condiciones en R | Usar la función *“unique”* para obtener los resultados únicos según la condición por la cual estemos filtrando información |
| III. Utilización de Scripts en R con aplicación al análisis de datos desde archivos | Utilizar un script para análisis de archivos |
| Tomar un dataset real y hacer preguntas sobre los datos |
| Utilizar funciones matemáticas en R |
| V. Estadística I. Análisis Exploratorio de datos: Análisis univariado y multivariado | Identificar si un dataframe tiene valores nulos |
| Encontrar las medidas de centralidad, dispersión y distribución de un dataframe |
| Identificar los tipos de datos de un dataframe y filtrar a categóricos |
| Construir una tabla de frecuencias |
| Encontrar relaciones entre variables |
| Crear visualizaciones |
| VI. Estadística II. Distribuciones muestral y paramétrica | Obtener muestras con o sin reemplazo de los renglones de un dataframe |
| Utilizar muestras para crear un dataframe nuevo solo con los índices que aparecieron en la muestra |
| Visualizar la distribución de todas las variables numéricas, tanto de una muestra como de una población |
|
| VII. Modelado Matemático: Regresión y Predicción | Implementar modelos de Regresión lineal en R |

# Bibliografía

* addieldealbs. (s.f.). *Módulo: Programación y Estadística con R*. Recuperado el 1 de Diciembre de 2020, de GitHub: <https://github.com/beduExpert/A2-Estadistica-Programacion-con-R-2020>
* CSSEGISandData. (s.f.). *csse\_covid\_19\_daily\_reports*. Recuperado el 4 de Diciembre de 2020, de GitHub: <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports>