Actividad 1 **Estadística Inferencial** **Ingeniería Industrial** **Profesora: Iliana Paternina Ortega**

Nombre Completo: BRAIKAN ANDRES PIÑA SALCEDO, STEVEN DAVID GÓMEZ FOLIACO Fecha: 22/08/2023

Observaciones:

- **Todo lo realizado debe ser justificado de manera clara y precisa, con argumentos estadísticos, sino no se le dará valoración.**
- **La actividad puede ser entregada en parejas**
- **Cualquier intento de plagio y/o fraude será sancionado con una clasificación de 0.0**

Análisis previo de datos (Grupo 1)

Parte 1. Análisis de valores atípicos

1. Para realizar la Actividad 1, debes crear un Rscript en RStudio y guardar el archivo con un nombre específico de tu preferencia.

2. Instala el paquete “carData” y carga los datos llamados “States”

3. Describe el dataset y cada una de sus variables
   Dentro el dataset de datos se encuentran 7 variables

Variable	Descripción	Tipo de dato
región	Regiones del censo de EE. UU. Un factor con niveles: ENC, Centro Norte Este; ESC, Centro Sur Este; MA, Atlántico Medio; MTN, Montaña; NE, Nueva Inglaterra; PAC, Pacífico; SA, Atlántico Sur; WNC, Centro-Noroeste; WSC, Centro Sur Oeste.	FACTOR
pop	Población: en miles.	INT
SATV	Puntaje promedio de los estudiantes que se gradúan de secundaria en el estado en el componente verbal de	INT

	la Prueba de Aptitud Escolar (un examen estándar de admisión a la universidad).	
SATM	Puntaje promedio de los estudiantes graduados de secundaria en el estado en el componente de matemáticas de la Prueba de Aptitud Escolar.	INT
percent	Porcentaje de estudiantes graduados de secundaria en el estado que tomaron el examen SAT.	INT
dollars	Gasto estatal en educación pública, en miles de dólares por estudiante.	NUM
pay	Salario promedio de un maestro en el estado, en miles de dólares.	INT

4. Realiza un summary de la variable dollars e interpreta dos estadísticas.
   Sabiendo que dentro de la variable Dollars se encuentran el gasto estatal en educación pública.

```
> summary(df$dollars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.993   4.354   5.045   5.175   5.689   9.159
```

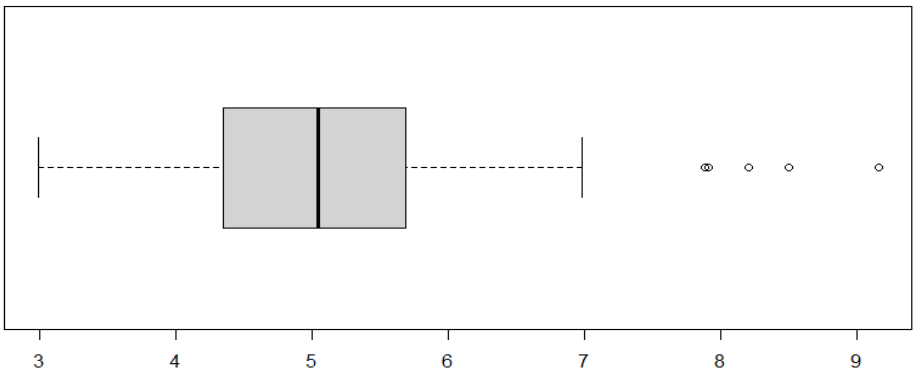
   Tomando en cuenta la *media y la mediana* se puede determinar que:

   En promedio por estudiante el gasto estatal en educación pública es de $5.175 mil dólares. mientras, mientras que la mediana determina qué valores están por encima de $5.045 mil dólares y la otra mitad está por debajo

5. Para la variable dollars realiza un análisis de datos atípicos y realiza la depuración completa de la variable. Debes ir copiando todos los resultados que vayas obteniendo en R e ir interpretando y/o explicando tus procedimientos.

   Boxplot número 1:Este boxplot demuestra la presencia de 5 datos atípicos.

```
boxplot(df$dollars, horizontal = TRUE)$out
which(df$dollars%in%boxplot(df$dollars)$out)
```



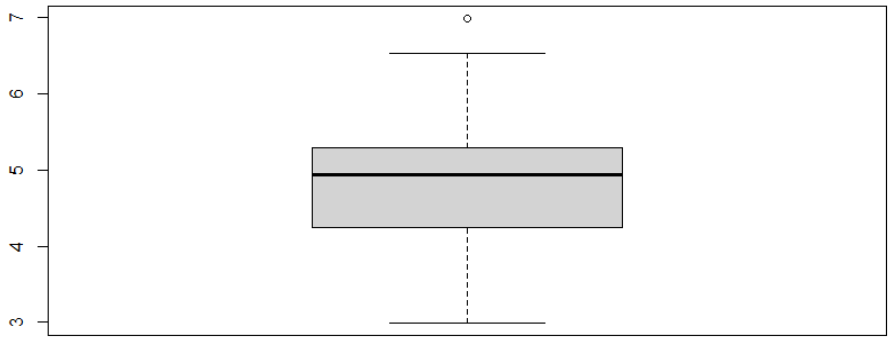

```
> which(df$dollars%in%boxplot(df$dollars)$out)
[1]  2  7  9 31 33
> dfDollar<-df$dollars[-c(2,7,9,31,33)]
> boxplot(dfDollar, horizontal = TRUE)$out
[1] 6.989
> which(dfDollar%in%boxplot(dfDollar)$out)
[1] 35
>
```

   Entonces, se decidió eliminar estos datos atípicos.

Por alguna razón quedó un dato atípico.

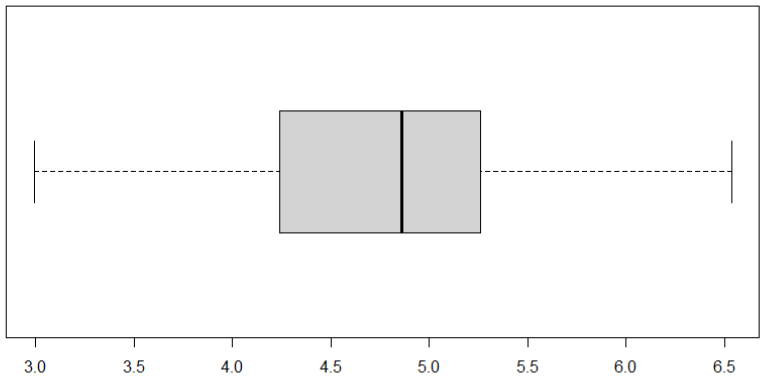
```
> boxplot(dfDollar, horizontal = TRUE)
> which(dfDollar%in%boxplot(dfDollar)$out)
[1] 35
> dfDollar <- dfDollar[-35]
> boxplot(dfDollar, horizontal = TRUE)
```

Boxplot número dos:



Al quitar ese dato atípico de la colección de datos se logra apreciar que ya no existen datos atípicos dentro de la colección de datos.

Boxplot número Tres:



6. Realiza un summary de la variable dollars e interpreta dos estadísticas (las mismas del punto 4) de cada variable y compara con los resultados del punto 4

   Nueva

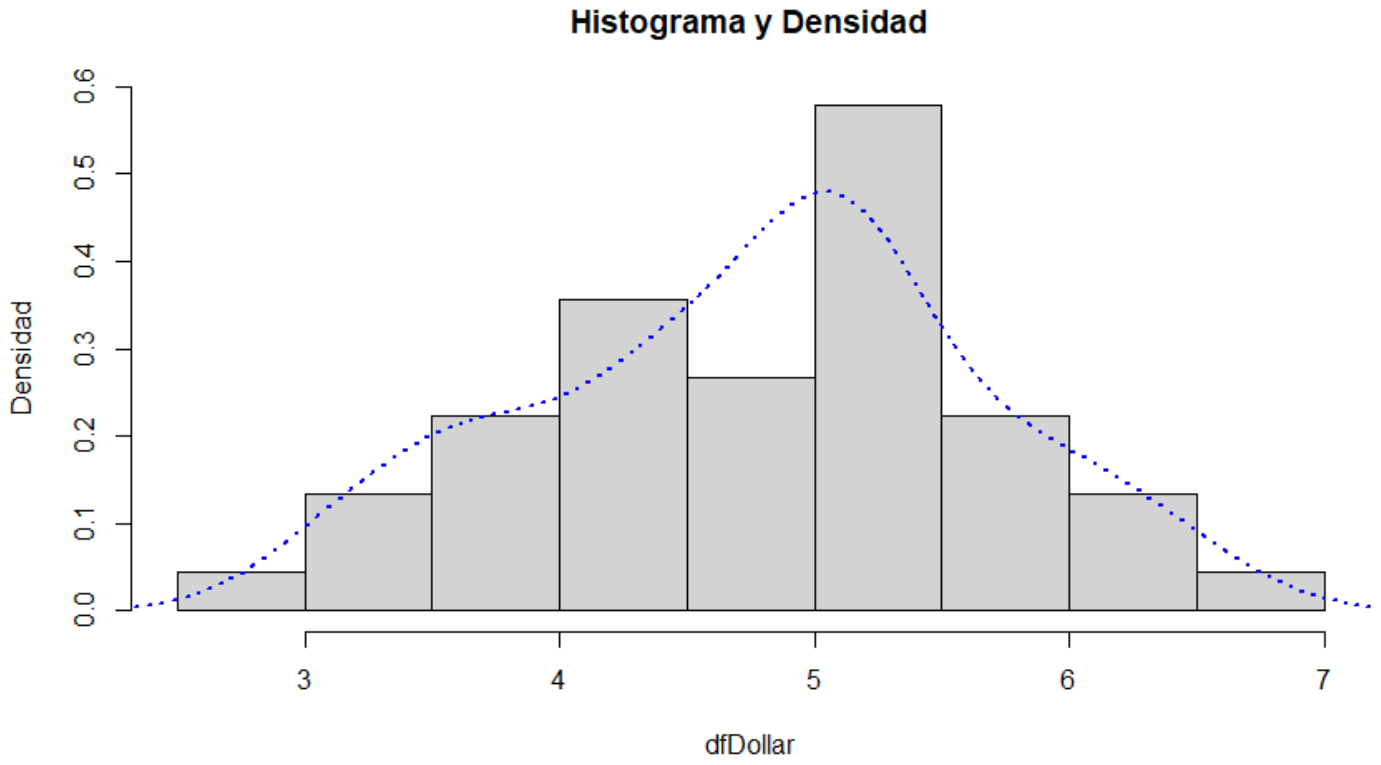
```
summary(dfDollar)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.993   4.238   4.860   4.784   5.260   6.534
```

Anterior

```
> summary(df$dollars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.993   4.354   5.045   5.175   5.689   9.159
```

Se determina que la tanto como la media y la mediana han disminuido significativamente debido a que se quitaron números superiores a la mediana.

7. Revisa el supuesto de normalidad para la variable dollars. Explica e interpreta los resultados que encuentras. Debes mostrar por lo menos un análisis gráfico y una prueba o test estadístico.



A Través de este gráfico de histograma y densidad se puede inferir que se parece a una distribución sesgada a la izquierda.

Se realizó el test de shapiro el cual logró determinar que esto significa que no hay suficiente evidencia para rechazar la hipótesis nula de que los datos de la variable siguen una distribución normal.

```
> shapiro.test(dfDollar)

        Shapiro-Wilk normality test

data:  dfDollar
W = 0.97894, p-value = 0.5773
```

Parte 2. Análisis de valores ausentes

1. Carga los datos llamados “attenu”

2. Describe el dataset y cada una de sus variables

Variable	Descripción	Tipo de dato
event	Número de evento	NUM
mag	Momento Magnitud	NUM
station	Número de estación	FACTOR
Dist	Distancia estación-hipocentro (km)	NUM
accel	Aceleración máxima (g)	NUM

3. Realiza un análisis de valores ausentes:
   - menciona si hay variables con valores ausentes,
   - describe la (s) variable(s) que hayas encontrado con valores ausentes,
   - realiza la imputación de la misma y realiza un summary e interpreta
   - Realiza una comparativa de las estadísticas que obtienes del summary de la variable omitiendo los valores ausentes y después de imputarlos.

La variable que tiene valores ausentes es “Station” la cual cuenta con 16 datos ausentes y que a lo largo de las siguientes instrucciones se imputarán estos datos.

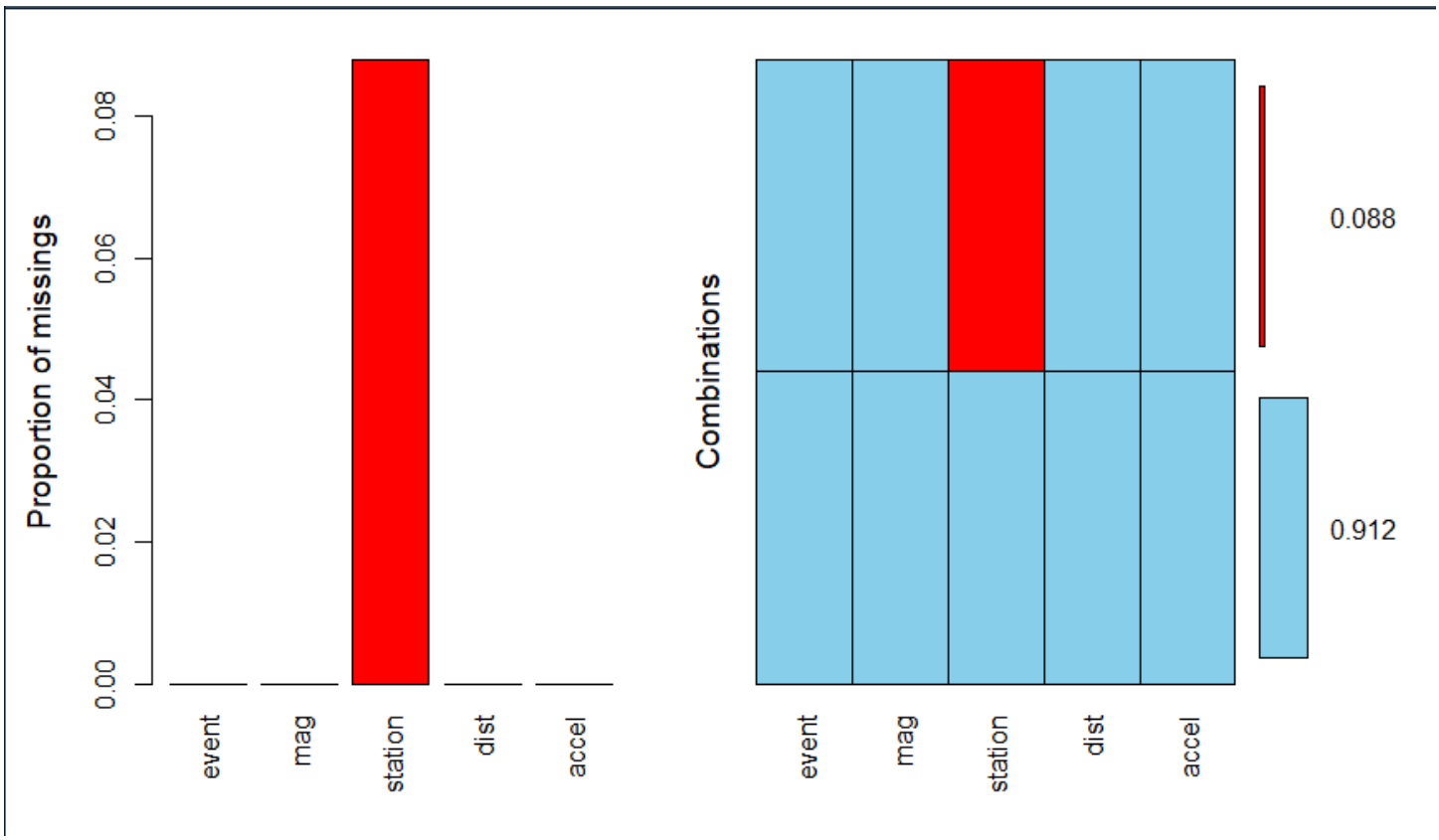
Imputando con el promedio

```
datos_imputados <- mice(attenu, method = "mean", print=FALSE)
datos_completos <- mice::complete(datos_imputados)
summary(datos_completos)
sum(is.na(datos_completos))

aggr(datos_completos, numbers=TRUE)
```

Aquí se muestran los datos con datos ausentes

```
> summary(datos_completos)
     event            mag            station        dist             accel
 Min.   : 1.00   Min.   :5.000   117     :  5   Min.   :  0.50   Min.   :0.00300
 1st Qu.: 9.00   1st Qu.:5.300   1028    :  4   1st Qu.: 11.32   1st Qu.:0.04425
 Median :18.00   Median :6.100   113     :  4   Median : 23.40   Median :0.11300
 Mean   :14.74   Mean   :6.084   112     :  3   Mean   : 45.60   Mean   :0.15422
 3rd Qu.:20.00   3rd Qu.:6.600   135     :  3   3rd Qu.: 47.55   3rd Qu.:0.21925
 Max.   :23.00   Max.   :7.700   (Other):147   Max.   :370.00   Max.   :0.81000
                                 NA's    : 16
> sum(is.na(datos_completos))
[1] 16
```



Imputamos con regresión.

```
#Imputar datos usando la regresion
datos_imputados2 <- mice(attenu, method = "norm.predict", print=FALSE)
datos_completos2 <- mice::complete(datos_imputados2)A
summary(datos_completos2)
sum(is.na(datos_completos2))
aggr(datos_completos2, numbers=TRUE)
```

```
> summary(datos_completos2)
     event            mag           station         dist            accel
 Min.   : 1.00   Min.   :5.000   117     :  5   Min.   :  0.50   Min.   :0.00300
 1st Qu.: 9.00   1st Qu.:5.300   1028    :  4   1st Qu.: 11.32   1st Qu.:0.04425
 Median :18.00   Median :6.100   113     :  4   Median : 23.40   Median :0.11300
 Mean   :14.74   Mean   :6.084   112     :  3   Mean   : 45.60   Mean   :0.15422
 3rd Qu.:20.00   3rd Qu.:6.600   135     :  3   3rd Qu.: 47.55   3rd Qu.:0.21925
 Max.   :23.00   Max.   :7.700   (Other):147   Max.   :370.00   Max.   :0.81000
                                 NA's    : 16
> sum(is.na(datos_completos2))
[1] 16
> aggr(datos_completos2, numbers=TRUE)
```

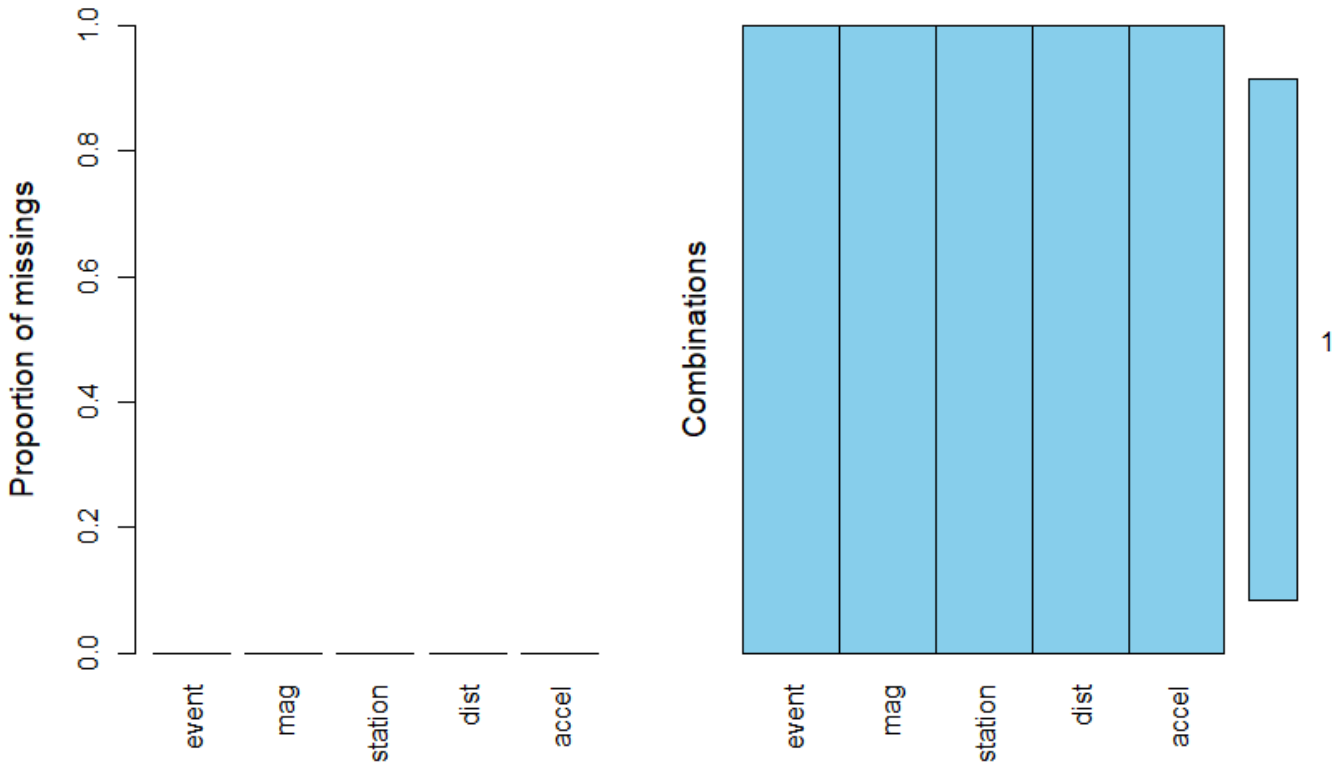
imputando usando el método pmm

```
#Imputar datos usando el metodo "pmm"
datos_imputados3 <- mice(attenu, method = "pmm", print=FALSE)
datos_completos3 <- mice::complete(datos_imputados3)
summary(datos_completos3)
sum(is.na(datos_completos3))
aggr(datos_completos3, numbers=TRUE)
```

Se quitaron los números atípicos

Nombre Completo: BRAIKAN ANDRES PIÑA SALCEDO, STEVEN DAVID GÓMEZ FOLIACO Fecha: 22/08/2023

```
> datos_imputados3 <- mice(attenu, method = "pmm", print=FALSE)
> datos_completos3 <- mice::complete(datos_imputados3)
> summary(datos_completos3)
     event             mag            station         dist             accel
 Min.   : 1.00   Min.   :5.000   117     :  5   Min.   :  0.50   Min.   :0.00300
 1st Qu.: 9.00   1st Qu.:5.300   1028    :  4   1st Qu.: 11.32   1st Qu.:0.04425
 Median :18.00   Median :6.100   113     :  4   Median : 23.40   Median :0.11300
 Mean   :14.74   Mean   :6.084   5055    :  4   Mean   : 45.60   Mean   :0.15422
 3rd Qu.:20.00   3rd Qu.:6.600   1030    :  3   3rd Qu.: 47.55   3rd Qu.:0.21925
 Max.   :23.00   Max.   :7.700   112     :  3   Max.   :370.00   Max.   :0.81000
                                 (Other):159
> sum(is.na(datos_completos3))
[1] 0
> aggr(datos_completos3, numbers=TRUE)
```



Podemos concluir que después de hacer la comparación entre los summary de los distintos métodos llegamos a imputar los NA.

SUBE A SAVI EL DOCUMENTO EN WORD CON TUS RESULTADOS E INTERPRETACIONES
SUBE A SAVI EL ARCHIVO DE R DONDE DEJASTE TU CÓDIGO