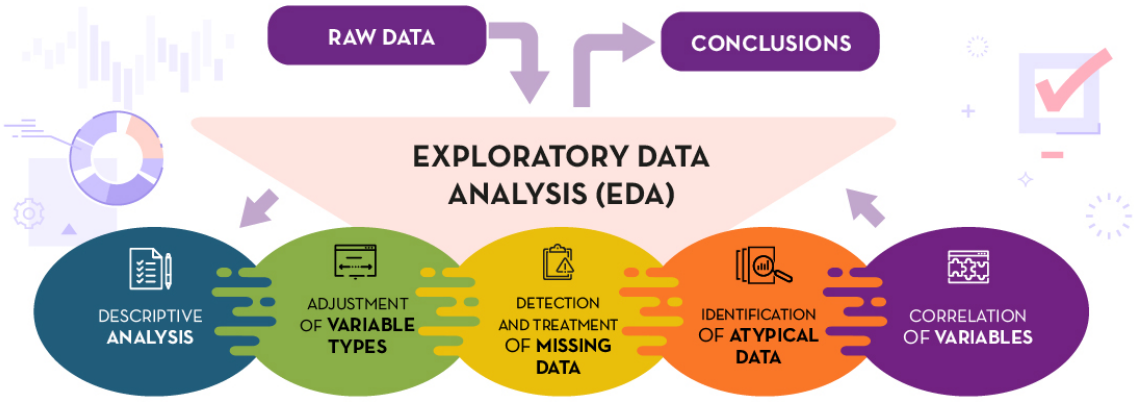
Millon de registros

Carlos Garcia Diaz, 1712253 Leonardo Cortez Gomez,

2023-09-08



1. Importando bibliotecas y el conjunto de datos

Bibliotecas necesarias para la visualizacion de los datos

```
library(ggplot2)
```

Se carga el archivo `Millon.csv`

```
data <- read.csv("Millon.csv")
```

2. Vista pevia de los datos

Se muestan los primeros 6 resgistros, con lo que podemos ver que el cojunto de datos consta de 5 columnas:

- Nombre
- Numero
- Promedio
- Semestre
- Materias

```r
head(data)
```

```
##           Nombre       Numero Promedio Semestre Materias
## 1   Jemima Berry 848-261-6134      7.0        5        7
## 2  Fritz Gardner 584-763-4957      6.6        5        2
## 3     Kevin Todd 183-377-3727      7.0        6       10
## 4 Rhiannon Kline 748-620-0328      6.8        7        3
## 5    Joan Monroe 895-282-3983      8.2        4        8
## 6 Libby Anderson 734-667-6043      7.7        7        6
```

Para conocer la estructura del conjunto de datos. Asi verificamos el tipo de datos de cada columna y validamos la integridad de estos.

```r
str(data)
```

```
## 'data.frame':    1000000 obs. of  5 variables:
##  $ Nombre  : chr  "Jemima Berry" "Fritz Gardner" "Kevin Todd" "Rhiannon Kline" ...
##  $ Numero  : chr  "848-261-6134" "584-763-4957" "183-377-3727" "748-620-0328" ...
##  $ Promedio: num  7 6.6 7 6.8 8.2 7.7 6.9 9 7.1 6.9 ...
##  $ Semestre: int  5 5 6 7 4 7 4 6 1 3 ...
##  $ Materias: int  7 2 10 3 8 6 10 7 2 5 ...
```

Se interpreta el *signinificado* de los datos y comprobamos que la columna:

- `Nombre` *del alumno* es de tipo char
- `Numero` *de telefono* es de tipo char, lo cual puede que no sea conveniente
- `Promeddio` *general* es de tipo num (variable continua)
- `Semestre`*s cursados* es de tipo entero
- `Materias` *que cursa en el semestre actual* es de tipo entero

3. Descripcion de los datos

Obtenemos un resumen del conjunto de datos:

```r
summary(data)
```

```
##     Nombre             Numero            Promedio        Semestre
##  Length:1000000     Length:1000000     Min.   : 6.000   Min.   :1.000
##  Class :character   Class :character   1st Qu.: 7.000   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median : 8.000   Median :4.000
##                                        Mean   : 7.999   Mean   :4.003
##                                        3rd Qu.: 9.000   3rd Qu.:6.000
##                                        Max.   :10.000   Max.   :7.000
##     Materias
##  Min.   : 2
##  1st Qu.: 4
##  Median : 6
##  Mean   : 6
##  3rd Qu.: 8
##  Max.   :10
```

De igual forma confirmamos que no hay datos perdidos (**NA**) en el conjunto de datos.

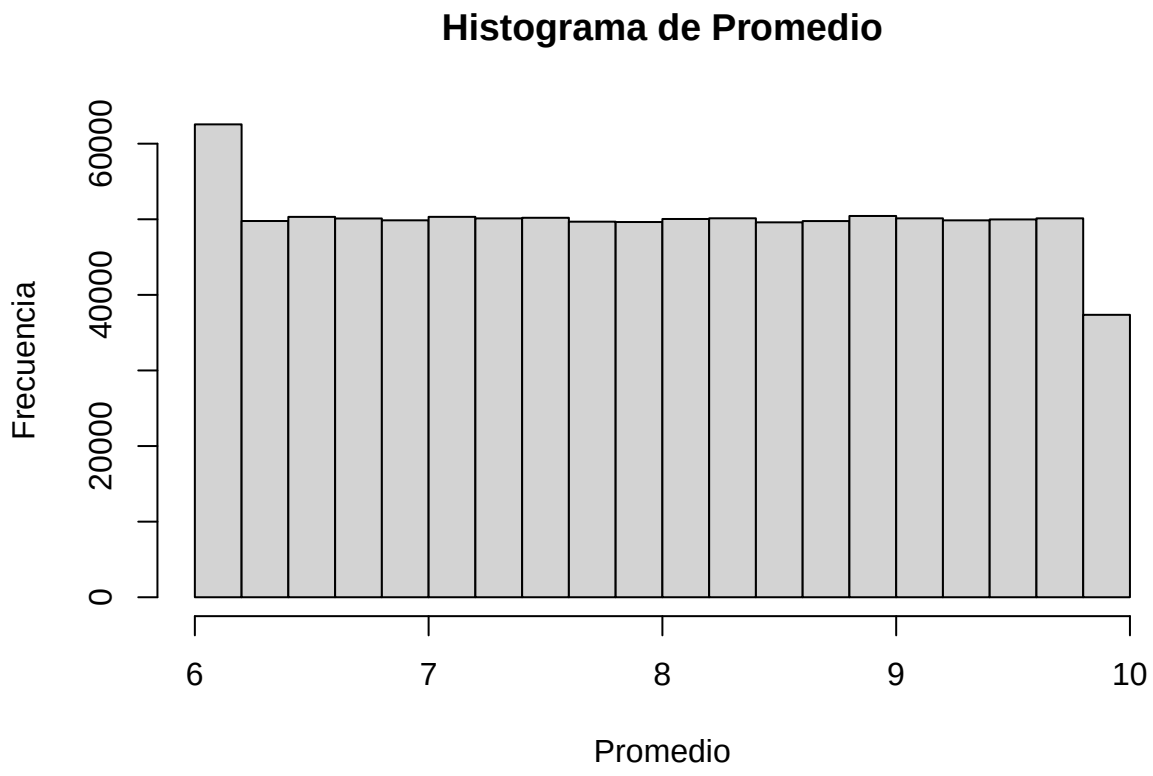
En cuanto a los datos numericos, vemos que:

- Para la columna `Promedio` (de todos los alumnos):
  - El valor minimo es de 6,0
  - El valor maximo es de 10.0
  - El promedio es de 7.99
- Para la columna `Semestre`:
  - El valor minimo es de 1
  - El valor maximo es de 7
  - El promedio es de 4.003, redondeado a 4
- Para la columna `Materias`:
  - El valor minimo es de 2
  - El valor maximo es de 10
  - El promedio es de 6

4. Visualizando los datos

Para tener una mejor comprension de los datos, se usan herramientas visuales como graficas

4.1 Promedio

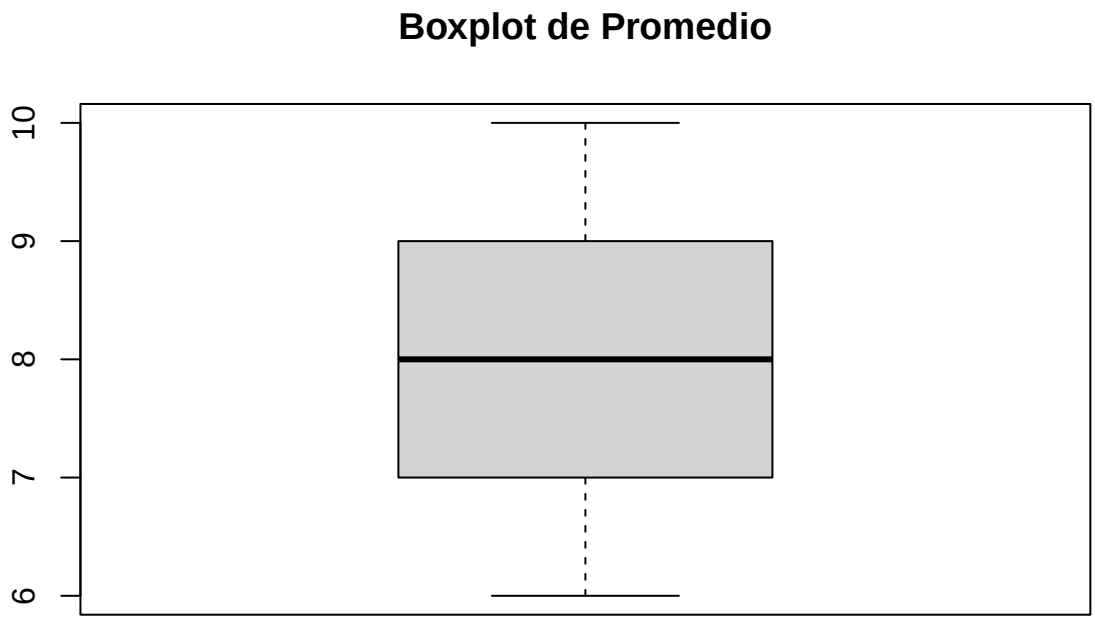


Teniendo en cuenta que cada intervalos es de 0.2, podemos observar quue mas de 60000 alumnos tienen un promedio menor o igual a 6.2. Mientras que menos de 40000 alumnos tienen un promedio mayor o igual a 9.8. Podemos confirmar esto calculando la desviación estándar de dichos datos:

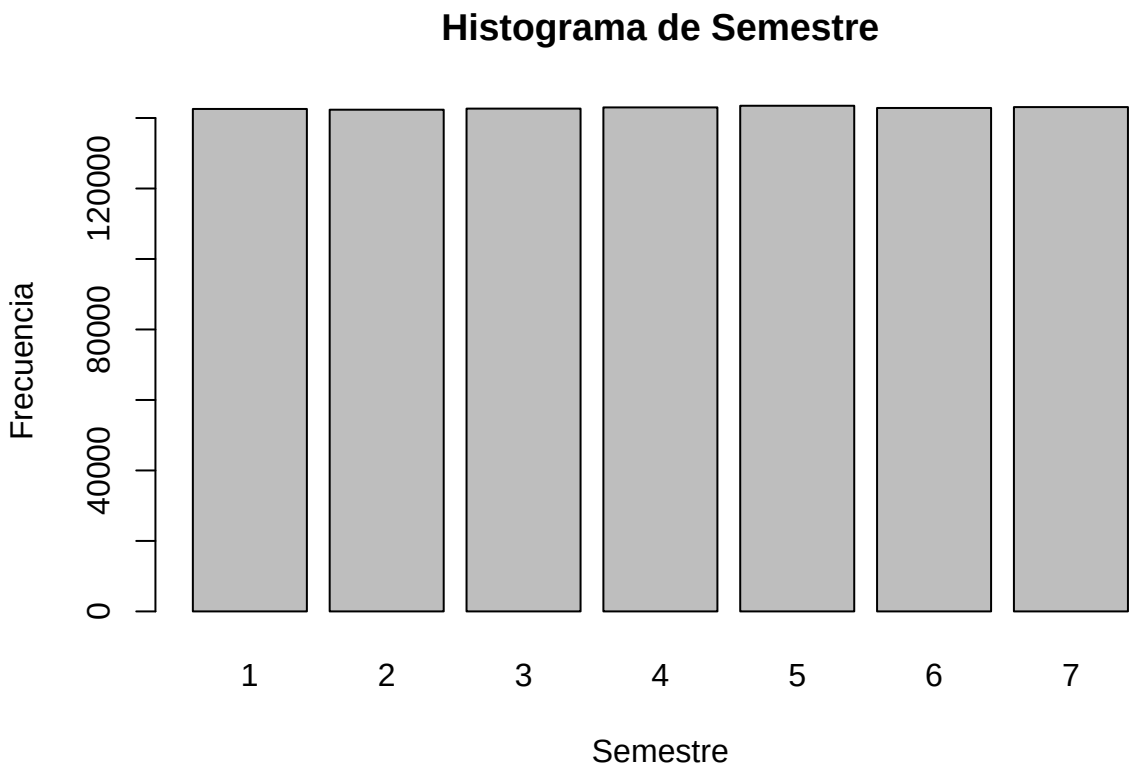
```
sd(data$Promedio)
```

```
## [1] 1.155638
```

Este resultado nos indica que, en promedio, los datos están alejados de la media en una valor de 1.15. También lo podemos obersvar con la siguiente gráfica.



4.2 Semestre

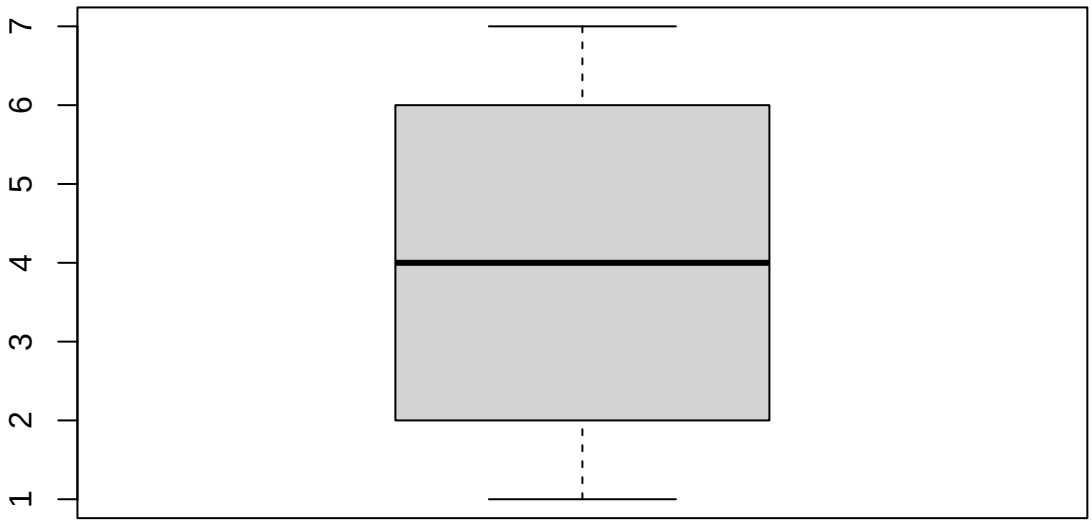


Con esta gráfica observamos que los valores de `Semestre` tienen una distribución normal, lo que indica que en los 7 semestres estudia casi la misma cantidad de alumnos. Podemos confirmar esto calculando la desviación estándar de dichos datos:

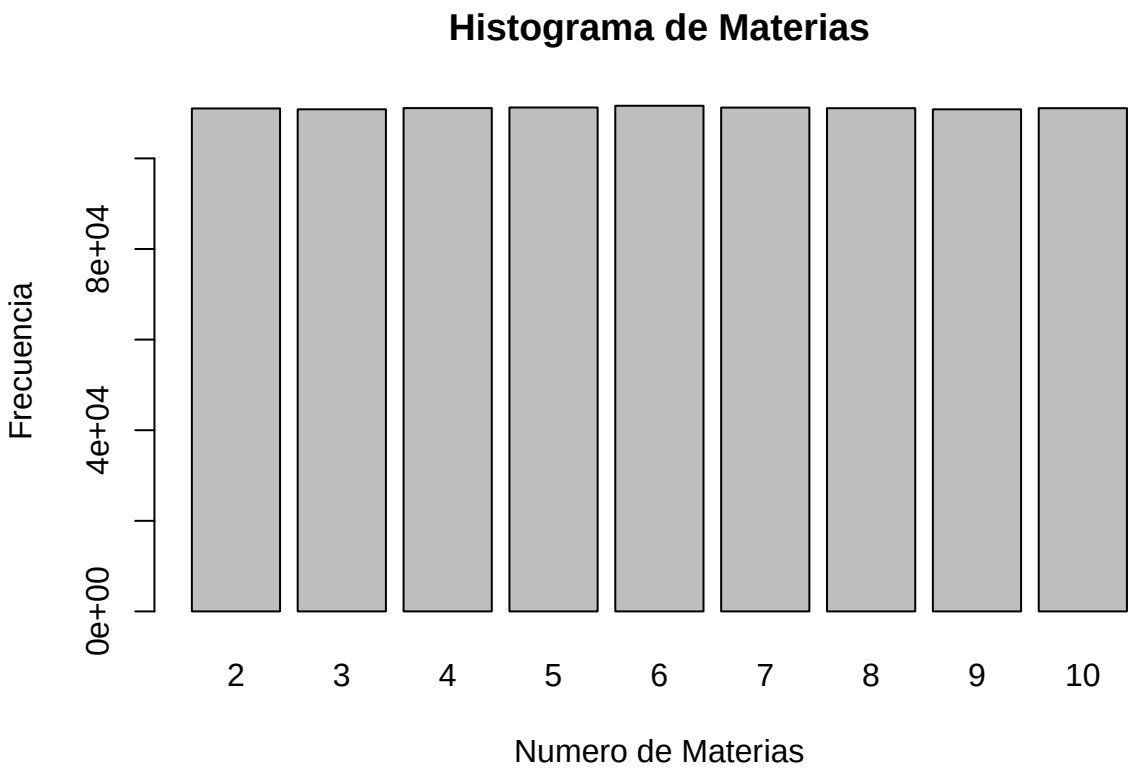
```
sd(data$Semestre)
```

```
## [1] 1.999453
```

Este resultado nos indica que, en promedio, los datos están alejados de la media en una valor de 1.99. También lo podemos obersvar con la siguiente gráfica.


Boxplot de Semestre
1
2
3
4
5
6
7

4.3 Materias



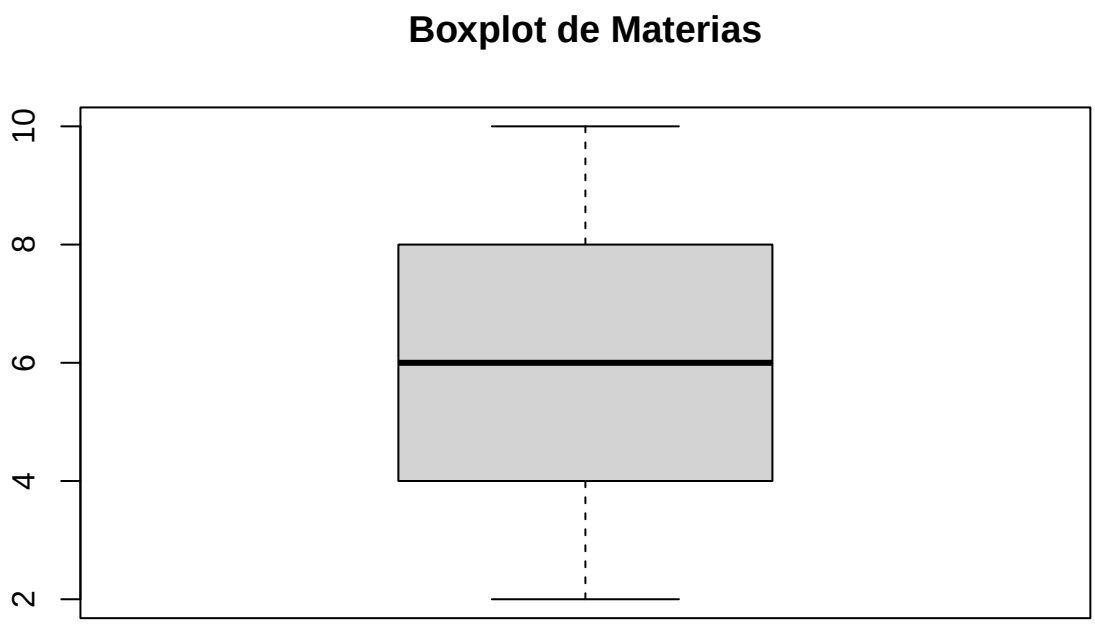
Asi como en la columna `Semestre`, aquí también podemos observar que el número de estudiantes que cursan de 1 a 10 materias es aproximadamente igual. Podemos confirmar esto calculando la desviación estándar de dichos datos:

```
sd(data$Materias)
```

```
## [1] 2.580635
```

Este resultado nos indica que, en promedio, los datos están alejados de la media en una valor de 2.5.

También lo podemos obersvar con la siguiente gráfica.



4.4 Numero

En cuanto a los datos de `Numero`, estos no se estudian ya que no son relevantes para el caso de estudio, pues estos son aleatorios.

5. Relaciones entre las variables

Se analizan las rerlaciones y qué tanto afectan las variables `Semestre` y `Materias` a la variable `Promedio`.

5.1 Matriz de correlación

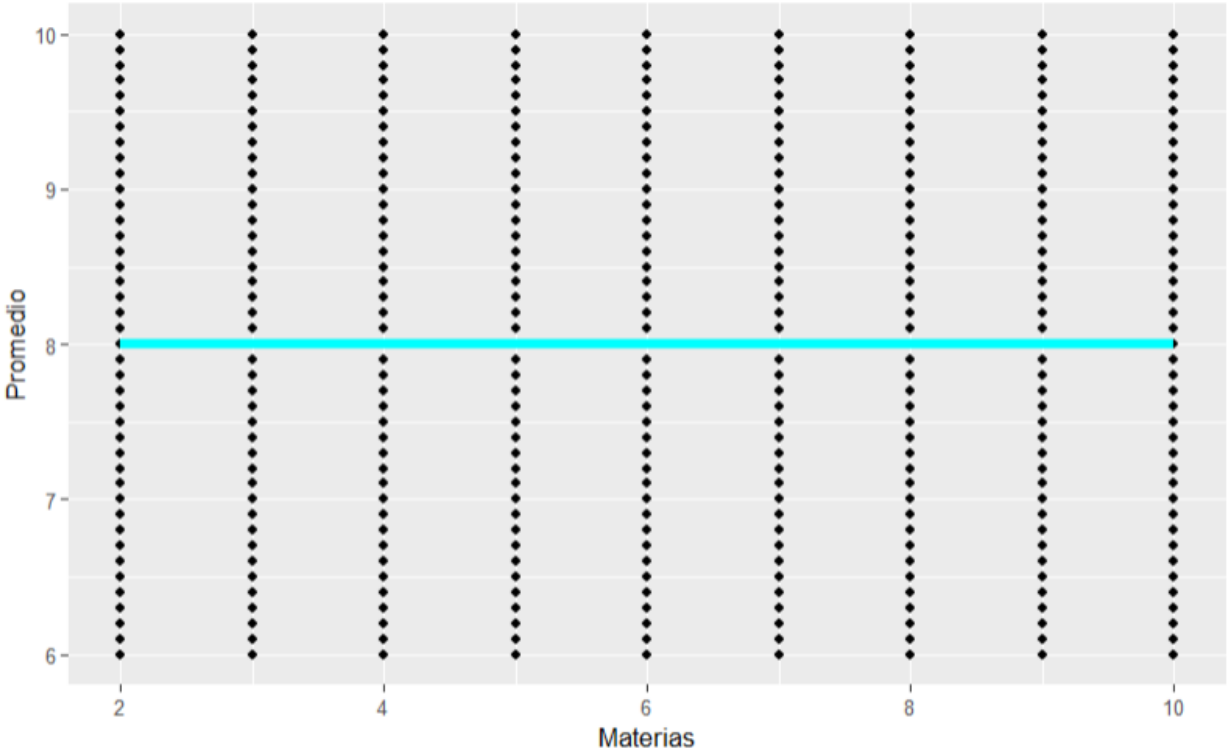
```
cor(data[,3:5])
```

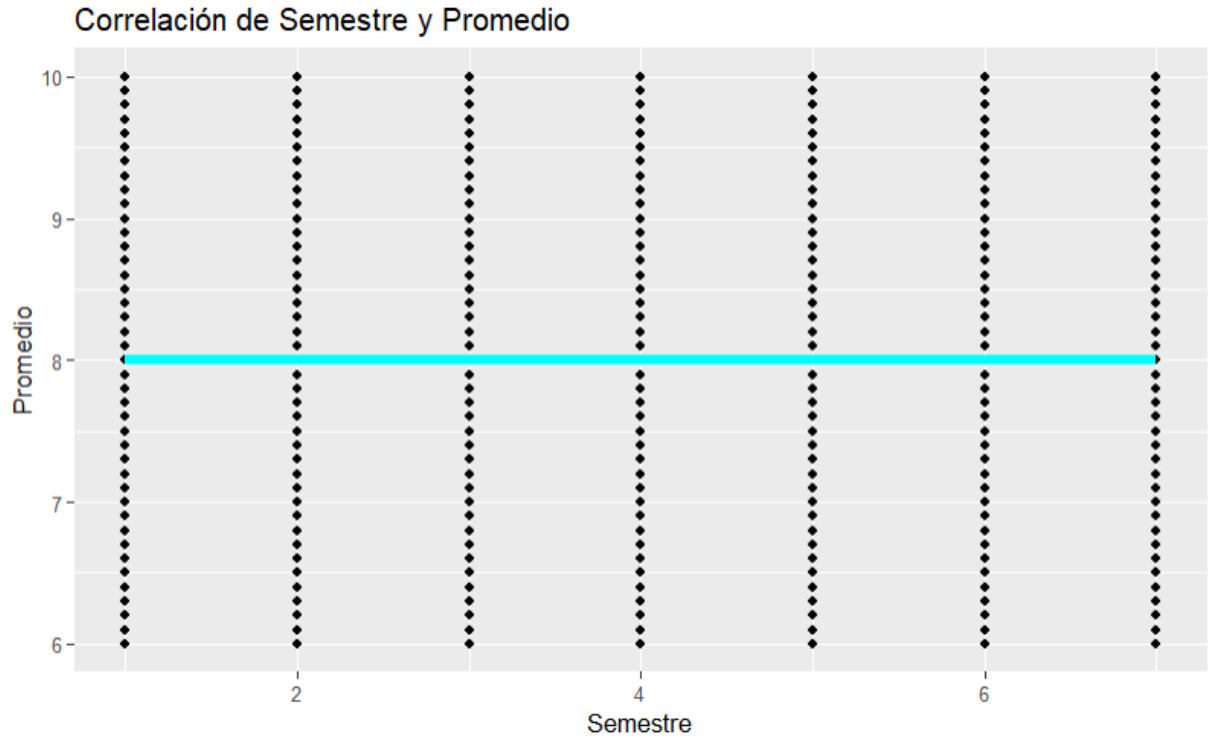
```
##              Promedio     Semestre     Materias
## Promedio 1.0000000000 0.0001444096 0.0001697354
## Semestre 0.0001444096 1.0000000000 0.0009683344
## Materias 0.0001697354 0.0009683344 1.0000000000
```

Dada la matriz de correlación, vemos que tanto el `Semestre` que cursa un alumno y el número de `Materias` que lleva en dicho semeste no afectan a su promedio pues los valores de correlación son prácticamente nulos.

5.2 Diagramas de dispersión

Para apreciar mejor dicha correlación, se presentan las siguientes gráficas junto con su linea de tendencia.





6. Conclusiones

A partir del análisis tanto gráfico como no gráfico describimos el comportamiento de las variables `Semestre` y `Promedio` y notamos que estas tienen una diespersión uniforme. Por otra parte, notamos que los valores de `Semestre` varían en sus extremos (6 y 10), pero en general conserva una distribución normal.

Y por último determinamos la correlación entre las variables para analizar cómo impactaban el valor del `Promedio`, con lo que concluimos que su impacto es mínimo.