Trabajo E-Portafolio

Grupo 6

2022-06-22

Introducción

El presente informe analiza una base de datos que contiene una muestra de 534 casos encuestados sobre salarios. Dichas encuestas fueron realizadas a un grupo de personas insertas en el mundo laboral, quienes fueron estudiados mediante 11 variables, definidas por:

- *Educación: Años de educación de la persona encuestada.*
- *Sector: Sector donde trabaja el encuestado. Puede trabajar en el sector privado (1) o en el sector público (0).*
- *Sexo: Masculino (0) o Femenino (1).*
- *Experiencia: Años de experiencia en el ámbito laboral.*
- *Edad: Edad del encuestado.*
- *Región: Región de en la cual habita la persona encuestada. Puede ser Metropolitana, Costa o Sur.*
- *Ocupación: Ocupación del encuestado. Puede ser en Transporte, Administración, Atención pública, Profesional, Servicios, Ventas u Otro.*
- *Área: Área en la cual trabaja el encuestado, puede ser en construcción, manufactura u otra.*
- *Casado: Corresponde a si el encuestado está Casado (1) o Soltero (0).*
- *Ingreso: Ingreso que percibe la persona encuestada en miles.*
- *Jornada: Jornada laboral. Puede ser completa (0) o parcial (1).*

El objetivo principal de este informe es analizar cada una de estas variables mediante los contenidos aprendidos en el curso como análisis descriptivos, modelos discretos y continuos. Con el fin de determinar posibles relaciones, estimaciones, observaciones anormales (outliers), sesgos, entre otros razonamientos que se llevarán a cabo en el desarrollo de este trabajo.

Considerando la situación actual, determinaremos cómo hoy en día difieren los salarios de las personas según su sexo, su educación, su ocupación y experiencia. Además, de analizar diversas variables para identificar las relaciones que puedan existir entre estas. Identificaremos si existe algún tipo de discriminación hacia un grupo determinado, la cual puede ser con respecto al sexo, experiencia, edad, educación, incluso estado civil de las personas encuestadas.

Análisis Descriptivo

En esta parte del trabajo, se presenta el análisis descriptivo realizado para cada variable de la muestra. Los resultados obtenidos se muestran a través de tablas de frecuencia, histogramas y distintos tipos de gráficos. Por otra parte, se realiza el análisis sobre la presencia de datos atípicos y tablas con medidas de posición y distribución.

Análisis de variables cuantitativas

En esta sección se analizará las variables cuantitativas entregadas en la base de datos.

Análisis de variable educación

Tabla de Frecuencia

educacion	frecuencia
4	1
6	5
7	5
8	18
9	11
10	22
11	27
12	225
13	37
14	52
15	15
16	65
17	22
18	29

Histograma

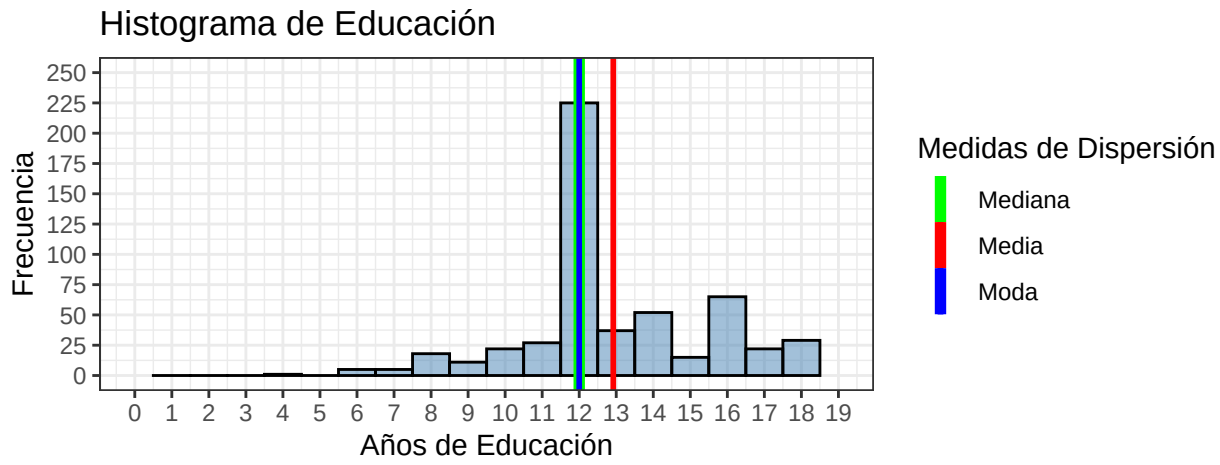


Tabla con medidas de dispersión y distribución

media	mediana	moda	varianza	desviacion	cv
12.91948	12	12	6.441909	2.538092	19.64547

Análisis de histograma y Tabla con medidas de dispersión y distribución

A partir de la tabla, se puede observar que la media es de *12.91948*, lo que implica que las personas estudiaron un promedio de aproximadamente 13 años. Si se ordenara los datos de manera ascendente, el promedio de los dos datos centrales, es de 12 años, lo que implica que esto es la mediana. El dato que más se repite es el que los encuestados estudian 12 años, siendo esta la moda.

Los resultados de estas tres medidas de localización y la interpretación del gráfico, muestra que los datos están sesgados positivamente con una asimetría a la derecha. Lo anterior se debe a que la media es mayor que la mediana y la moda.

En la tabla, tambien se obtuvieron las medidas de dispersión que corresponden a una varianza de *6.441909*, una desviación estandar de *2.538092* y un coeficiente de variación de *19.64547*. La desviación estandar indica que,la dispersión de los años de educación entre estas personas es de 2.5 años.

Medidas de posición

Rango

```
14
```

Cuartiles

```
  0%  25%  50%  75% 100% 
   4   12   12   14   18 
```

Rango intercuartilico

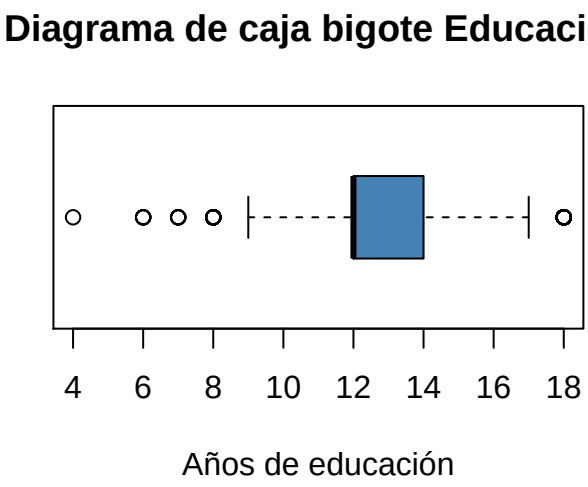
```
2
```

El rango representa la diferencia entre el año máximo que estudió un encuestado y el año mínimo. Esta diferencia corresponde a *14* años. Con base en los cuartiles, el 25% de los encuestados estudió entre 4 a 12 años. Por otra parte, el 50% estudió 12 años o menos. El 75% estudió 14 años o menos, y finalmente, el 100% de los encuestados estudió 18 años o menos.

El rango intercuartílico representa la distancia entre el cuartil 3 y el cuartil 1, donde se concentra el 50% de los datos. En este caso es de 2.

Estudio de datos atípicos

Para analizar la existencia de datos atípicos, se utilizó diagrama de caja bigote. Los datos atípicos corresponden a todos los datos que están a la izquierda del extremo inferior del diagrama y a la derecha del extremo superior del diagrama.



Analizando el diagrama de caja bigote, se puede observar que existen datos outliers, los cuales están ubicados en los valores 4,6,7,8 y 18 años de educación.

Análisis de variable experiencia

Histograma

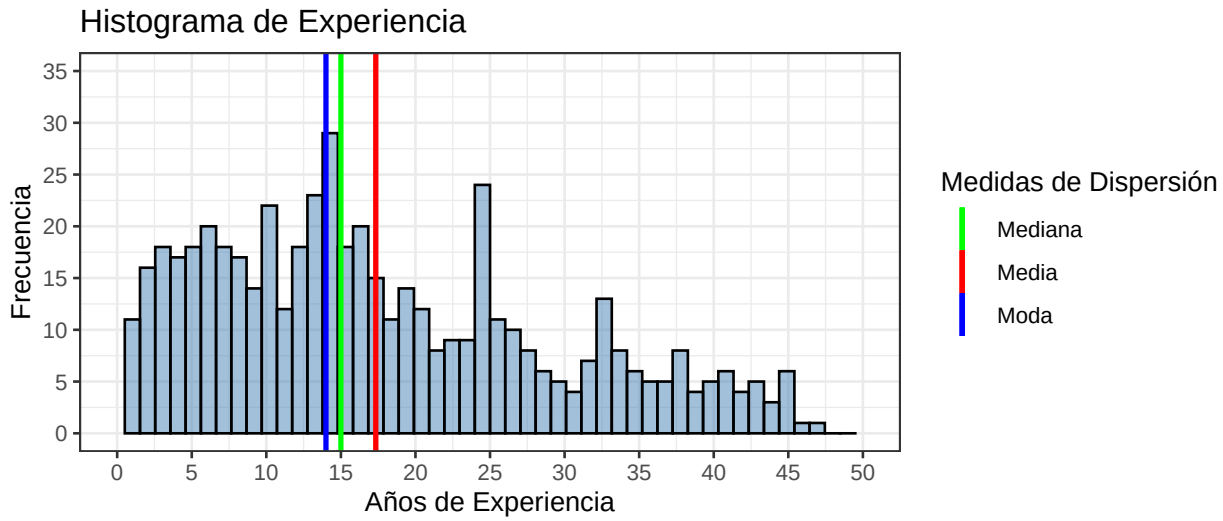


Tabla con medidas de dispersión y distribución

media	mediana	moda	varianza	desviacion	cv
17.33895	15	14	138.8886	11.7851	67.96895

Análisis de histograma y Tabla con medidas de dispersión y distribución

A partir de la tabla, se puede observar que la media es de *17.33895*, lo que implica que las personas tienen una experiencia en promedio de aproximadamente 17 años. Si se ordenara los datos de manera ascendente, el promedio de los dos datos centrales, es de 15 años, lo que implica que la mediana de estos datos son 15 años. El dato que más se repite es el que los encuestados tienen 14 años de experiencia, siendo esta la moda.

Los resultados de estas tres medidas de localización y la interpretación del gráfico, muestra que los datos están sesgados positivamente con una asimetría a la derecha. Lo anterior se debe a que la media es mayor que la mediana y la moda.

En la tabla, tambien se obtuvieron las medidas de dispersión que corresponden a una varianza de *138.8886*, una desviación estandar de *11.7851* y un coeficiente de variación de *67.96895*. La desviación estandar indica que,la dispersión de los años de experiencia entre estas personas es de 11.8 años aproximadamente.

Medidas de posición

Rango

```
47
```

Cuartiles

```
  0%  25%  50%  75% 100% 
   0    8   15   25   47 
```

Rango intercuartilico

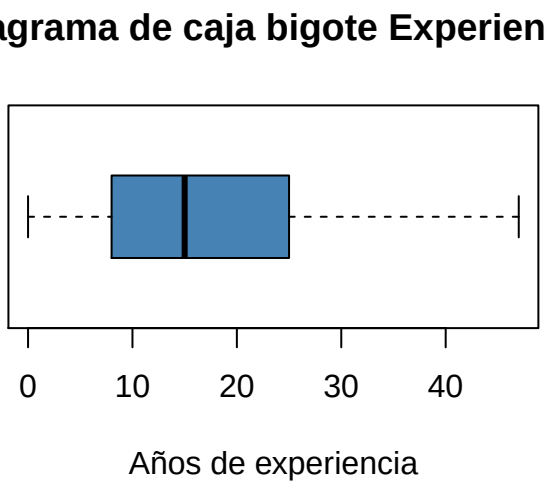
```
17
```

El rango representa la diferencia entre el año máximo de experiencia de un encuestado y el año mínimo. Esta diferencia corresponde a 47 años. Con base en los cuartiles, el 25% de los encuestados tiene una experiencia de 8 años o menos. Por otra parte, el 50% una experiencia de 15 años o menos. El 75% una experiencia de 25 años o menos, y finalmente, el 100% de los encuestados tiene una experiencia de 47 años o menos.

El rango intercuartílico representa la distancia entre el cuartil 3 y el cuartil 1, donde se concentra el 50% de los datos. En este caso es de 17.

Estudio de datos atípicos

Se utilizó el mismo metodo para analizar los datos atípicos que en la primera variable.



Analizando el diagrama de caja bigote, se puede observar que no existen datos outliers.

Análisis de variable edad

Histograma

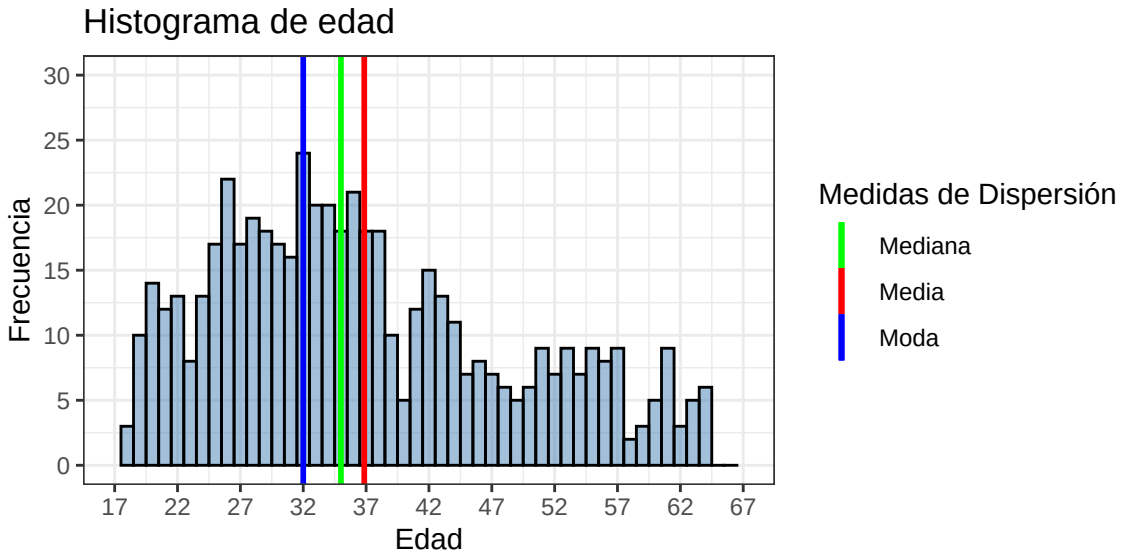


Tabla con medidas de dispersión y distribución

media	mediana	moda	varianza	desviacion	cv
36.85206	35	32	136.9931	11.7044	31.76052

Análisis de histograma y Tabla con medidas de dispersión y distribución

A partir de la tabla, se puede observar que la media es de *36.85206*, lo que implica que las personas encuestadas tienen una edad promedio de aproximadamente 37 años. Si se ordenara los datos de manera ascendente, el promedio de los dos datos centrales, es de 35 años, lo que implica que la mediana de estos datos son 35 años de edad. El dato que más se repite es el que los encuestados tienen 32 años de edad, siendo este la moda.

Los resultados de estas tres medidas de localización y la interpretación del gráfico, muestra que los datos están sesgados positivamente con una asimetría a la derecha. Lo anterior se debe a que la media es mayor que la mediana y la moda.

En la tabla, tambien se obtuvieron las medidas de dispersión que corresponden a una varianza de *136.9931* , una desviación estandar de *11.7044* y un coeficiente de variación de *31.76052*. La desviación estandar indica que,la dispersión de los años de edad entre estas personas es de 12 años.

Medidas de posición

Rango

```
46
```

Cuartiles

```
  0%  25%  50%  75% 100% 
  18   28   35   44   64 
```

Rango intercuartilico

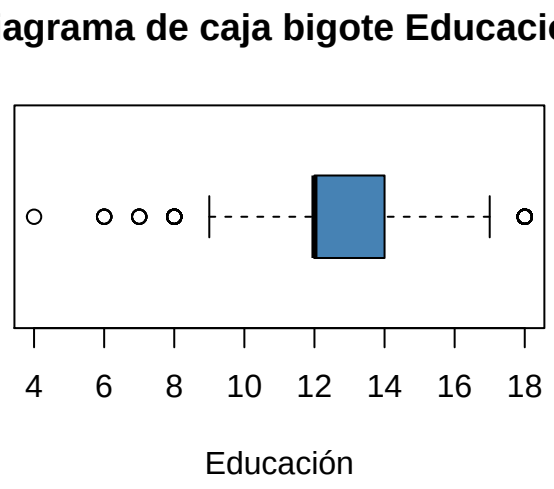
```
16
```

El rango representa la diferencia entre el la edad máxima de un encuestado y la edad mínima. Esta diferencia corresponde a *46* años. Con base en los cuartiles, el 25% de los encuestados tiene una edad entre 18 a 28 años. Por otra parte, el 50% tiene una edad de 35 años o menos. El 75% tiene una edad de 44 años o menos, y finalmente, el 100% de los encuestados tiene una edad de 64 años o menos.

El rango intercuartílico representa la distancia entre el cuartil 3 y el cuartil 1, donde se concentra el 50% de los datos. En este caso es de 2.

Estudio de datos atípicos

Se utilizó el mismo metodo para analizar los datos atípicos que en la primera variable.



Analizando el diagrama de caja bigote, se puede observar que existen datos outliers, los cuales están ubicados en los valores 4,6,7,8 y 18 años de educación.

Análisis de variable ingreso

Histograma

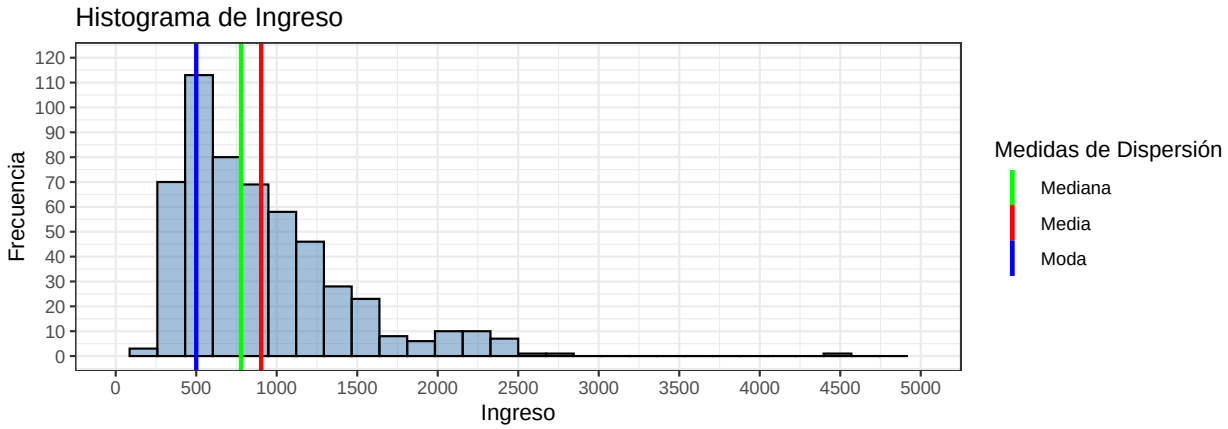


Tabla con medidas de dispersión y distribución

media	mediana	moda	varianza	desviacion	cv
903.2022	778	500	264375.9	514.175	56.928

Análisis de histograma y Tabla con medidas de dispersión y distribución

A partir de la tabla, se puede observar que la media de ingreso es de *903.2022*, lo que implica que las personas tienen un ingreso promedio de aproximadamente $903 000. Si se ordenara los datos de manera ascendente, el promedio de los dos datos centrales, es de 778, lo que implica que la mediana de los ingresos es de $778 000. El dato que más se repite es el que los encuestados tienen un ingreso de $500 000 siendo este el ingreso de moda.

Los resultados de estas tres medidas de localización y la interpretación del gráfico, muestra que los datos están sesgados positivamente con una asimetría a la derecha. Lo anterior se debe a que la media es mayor que la mediana y la moda.

En la tabla, tambien se obtuvieron las medidas de dispersión que corresponden a una varianza de *264375.9* , una desviación estandar de *514.175* y un coeficiente de variación de *56.928*. La desviación estandar indica que,la dispersión de la cantidad de ingreso entre estas personas es de $514 175.

Medidas de posición

Rango

```
4325
```

Cuartiles

```
  0%  25%  50%  75% 100% 
 125  525  778 1125 4450 
```

Rango intercuartilico

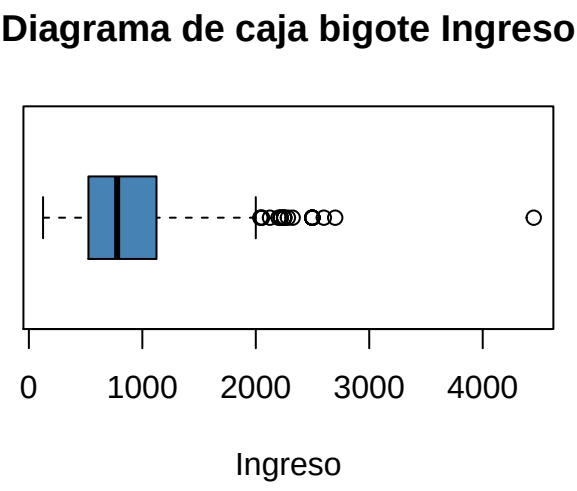
```
600
```

El rango representa la diferencia entre el ingreso máximo que gana un encuestado y el ingreso mínimo. Esta diferencia corresponde a *$4 325 000*. Con base en los cuartiles, el 25% de los encuestados tiene un ingreso entre $125 000 a $525 000. Por otra parte, el 50% tiene un ingreso de $778 000 o menos. El 75% tiene un ingreso de $1 125 000 o menos, y finalmente, el 100% de los encuestados tiene un ingreso de $4 450 000 o menos.

El rango intercuartílico representa la distancia entre el cuartil 3 y el cuartil 1, donde se concentra el 50% de los datos. En este caso es de 600.

Estudio de datos atípicos

Para analizar la existencia de datos atípicos, se utilizó diagrama de caja bigote.



Analizando el diagrama de caja bigote, se puede observar que existen datos outliers, lo cuales se encuentran entre $2 000 000 y $ 5 000 000

Análisis de datos binarios

Los datos binarios entregados por la base de datos, son aquellas variables las cuales tinen 0 y 1. Sin embargo, se creó una nueva base de datos para darle valores a estos datos binarios.

Análisis de variable sector

sector	frecuencia
privado	156
publico	378

De la tabla, se observa que en el sector privado trabajan 156 personas. Por otra parte, en el sector público, trabajan 378 personas.

Modelo binomial

Para calcular la media, la varianza y la desviación estandar se utilizará este modelo

X~B(n,p)
X(La persona que trabaje en el sector privado será nuestro éxito)
p = Probabilidad de trabajar en el sector privado
n = Número de intentos = 534

La probalibidad de que las personas encuestadas trabajen en el sector privado es de *0.2921348*

Se espera que *156* personas trabajen en el sector privado

Varianza

```
110.427
```

Desviación estandar

```
10.50842
```

X~B(n,p)
X(La persona que trabaje en el sector público será nuestro éxito)
p = Probabilidad de trabajar en el sector privado
n = Número de intentos = 534

La probalibidad de que las personas encuestadas trabajen en el sector público es de *0.7078652*

Se espera que *378* personas trabajen en el sector público

Varianza

```
110.427
```

Desviación estandar

```
10.50842
```

Análisis de variable sexo

sexo	frecuencia
femenino	245
masculino	289

De la tabla, se observa que del total de los encuestados, 245 corresponden a mujeres. Mientras que 289 corresponden a hombres.

Modelo binomial

Para calcular la media, la varianza y la desviación se utilizará este modelo

X~B(n,p)
X(La persona encuestada, que su sexo sea mujer, será nuestro éxito)
p = Probabilidad de ser mujer
n = Número de intentos = 534

Del total de personas encuestadas, la probabilidad de ser mujer es de *0.4588015.*

Se espera que, de las personas encuestadas, *245* sean mujeres.

Varianza

```
132.5936
```

Desviación estándar

```
11.51493
```

X~B(n,p)
X(La persona encuestada, que su sexo sea hombre, será nuestro éxito)
p = Probabilidad de ser hombre
n = Número de intento = 534

Del total de personas encuestadas, la probabilidad de ser hombre es de *0.5411985*

Se espera que, de las personas encuestadas, *289* sean hombres.

Varianza

```
132.5936
```

Desviación estándar

```
11.51493
```

Análisis de variable casado

casado	frecuencia
casado	350
soltero	184

De la tabla, se observa que 350 personas están casadas y 184 personas se encuentran solteras.

Modelo binomial
Para calcular la media, la varianza y la desviación se utilizará este modelo

X~B(n,p)
X(El éxito será que la persona encuestada esté casada)
p = Probabilidad de que la persona encuestada esté casada
n = Número de intento = 534

La probabilidad de que alguna persona de las encuestadas esté casada es de *0.6554307*

Se espera que *350* personas estén casadas

Varianza

```
120.5993
```

Desviación estándar

```
10.98177
```

X~B(n,p)
X(El éxito será que la persona encuestada esté soltera)
p = Probabilidad de que la persona encuestada esté soltera
n = Número de intentos = 534

La probabilidad de que un encuestado esté soltero o soltera es de *0.3445693.*

Se espera que *184* encuestados estén solteros o solteras.

Varianza

```
120.5993
```

Desviación estándar

```
10.98177
```

Análisis de variable jornada

jornada	frecuencia
completa	479
parcial	55

De la tabla, se observa que 479 personas trabajan en jornada completa y 55 personas trabajan en jornada parcial.

Modelo binomial

Para calcular la media, la varianza y la desviación se utilizará este modelo
X~B(n,p)
X(Si la persona encuestada trabaja en jornada completa se considerará éxito)
p = Probabilidad de que trabaje jornada completa
n = Número de intentos = 534

La probabilidad de que algún encuestado trabaje jornada completa es de *0.8970037.*

Se espera que *479* personas trabajen en jornadas completas
Varianza

```
49.33521
```

Desviación estándar

```
7.023902
```

X~B(n,p)
X(Si la persona encuestada trabaja en jornada parcial se considerará éxito)
p = Probabilidad de que trabaje jornada parcial
n = Número de intentos = 534

La probabilidad de que al encuestar una persona y que esta trabaje jornada parcial es de *0.1029963.*

Se espera que *55* personas trabajen en jornada parcial.

Varianza

```
49.33521
```

Desviación estándar

```
7.023902
```

Análisis de variable cualitativas

En las variables cualitativas solo podemos realizar tablas de frecuencia, ya que sus datos son valores de texto.

Análisis de variable región

region	frecuencia
Costa	27
Metrop	440
Sur	67

De la tabla se observa que 27 personas se encuentran en la región costa, 440 personas se encuentran en la región metropolitana y 67 personas se encuentran en la región sur.

Análisis de variable ocupación

ocupacion	frecuencia
Administrac	46
AtencionPub	77
Otro	2
Profesional	93
Servicios	60
Transporte	222
Ventas	34

De la tabla se observa que la ocupación de 46 personas es en administración, la de 77 es en atención pública, la de 93 personas es profesional, la de 60 personas es de servicios, la de 222 es de transporte, la de 34 personas son las ventas y 2 personas tienen otra ocupación.

Análisis de variable área

area	frecuencia
Construcion	24
Manufactura	98
Otro	412

De la tabla se observa que el área de trabajo de 24 personas es la construcción, la de 98 personas es manufactura y la 412 personas su área de trabajo es otra.

Medidas de asociación

Con respecto a la variable ingreso en datos binarios y varibles cualitativas.

Para las variables binarias y cualitativas se armaron tablas con los promedios de ingreso de cada una de las variables.

Variable sexo

sexo	promedio
femenino	789.2082
masculino	999.8408

En esta tabla se puede observar que en promedio, los hombres ganan más que las mujeres.

Variable sector

sector	promedio
privado	791.0320
publico	949.4947

Se observa que en promedio las personas que trabajan en el sector público ganan mas que los que trabajan en el sector privado.

Variable región

region	promedio
Costa	728.3333
Metrop	928.6159
Sur	806.7761

En la tabla se observa que las personas de la Región Metropolitana tienen en promedio un ingreso mayor a las otras regiones. La segunda región con mayores ingresos es la del Sur. Finalmente, la región con menos ingresos es la de la Costa.

Variable ocupación

ocupacion	promedio
Administrac	1326.9130
AtencionPub	759.0519
Otro	310.0000
Profesional	1206.4839
Servicios	592.7000
Transporte	844.5315
Ventas	792.7647

Se observa que, las personas que se ocupan en administración, en promedio, son las que más ingresos preciben. Mientras que las que menos ingresos perciben son las personas que tienen otra ocupación.

Variable área

area	promedio
Construcion	922.9167
Manufactura	966.4694
Otro	887.0049

Se observa que, en promedio, el área que más ingresos obtiene es la de manufactura. Después la de construcción y finalmente, las otras áreas.

Variable casado

casado	promedio
casado	940.9486
soltero	831.4022

Se observa que, en promedio, las personas casadas perciben más ingresos.

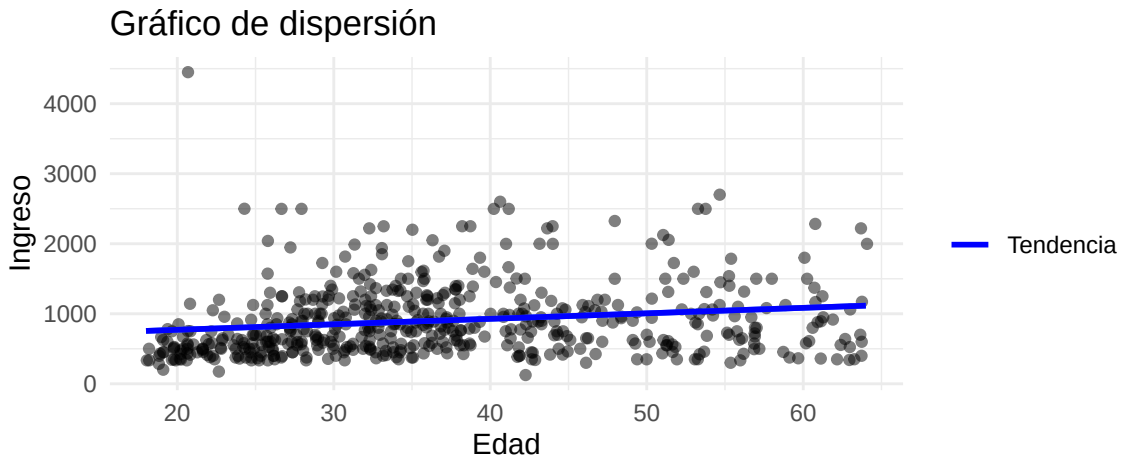
Variable jornada

jornada	promedio
completa	957.5511
parcial	429.8727

Aunque parezca una obviedad, las personas que trabajan en jornada completa, ganan más en comparación a los que trabajan en jornada parcial.

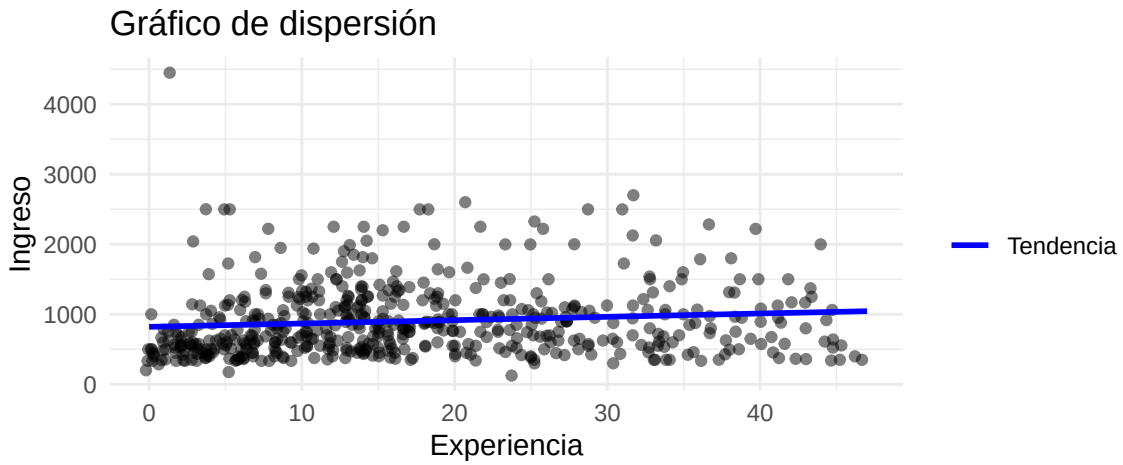
Con respecto a la variable ingreso y varibles cuantitativas.
En esta sección del informe, se realizó graficos de puntos con línea de tendencia para identificar el tipo de relación entre la variable ingreso y las otras variables cuantitativas. Además, se calculó la correlación para identificar con cual tiene una relacción positiva más fuerte.

Variable edad



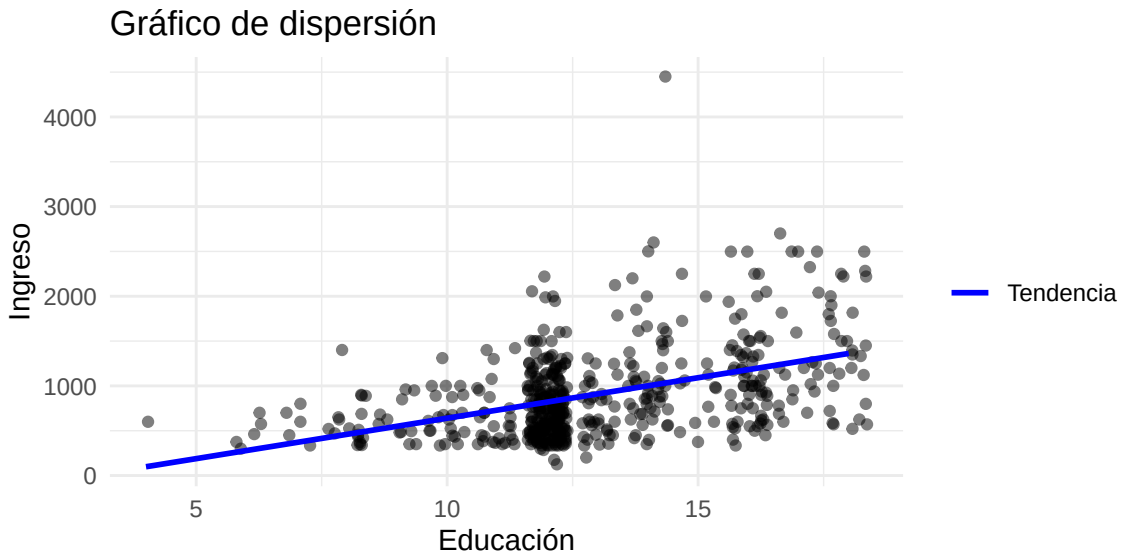
Se observa que a medida que aumenta la edad, la tendencia es que el ingreso percibido aumente para las personas.

Variable experiencia



Se observa que a medida que aumentan los años de experiencia de las personas, la tendencia es que su ingreso percibido aumente.

Variable educación



Se observa que a medida que aumentan los años de educación de las personas, la tendencia es que su ingreso percibido aumente.

Análisis de correlaciones

Correlación de ingreso con edad

```
0.1783684
```

Correlación de ingreso con experiencia

```
0.1090531
```

Correlación de ingreso con educación

```
0.445523
```

Se observa que la correlación más fuerte es la de ingreso con educación, por ende, cuando aumenta la variable educación, el ingreso aumenta en mayor medida que cuando aumenta las variables experiencia y edad.

Conclusiones

Luego de realizar un análisis descriptivo de cada variable e identificar la relación de cada una de estas con el **ingreso**, se puede concluir que:

Existe una brecha salarial entre los sexos. Dentro del análisis realizado, se observa que los hombres ganan más ingresos que las mujeres en promedio. Por lo tanto, tal como se dijo al inicio de este informe, se puede identificar que sí existe algún tipo de discriminación respecto al sexo de las personas.

Para percibir un salario alto es más importante los años de estudio de la persona, que su edad y los años que tiene de experiencia. Al analizar las correlaciones entre años de estudio, edad y años de experiencia, con respecto a salario, se observa que la correlación de años de estudio e ingreso posee el mayor valor, por lo que a medida que aumentan los años de estudio, los ingresos también tienden a aumentar en mayor medida de lo que aumenta la edad y los años de experiencia.

El sector público genera más ingresos. Se puede observar que en promedio las personas que trabajan en el sector público perciben más ingresos que las personas que trabajan en el sector privado.

Existe una brecha salarial entre las personas que viven en distintas regiones. Las personas que pertenecen a la Región Metropolitana ganan más ingresos en promedio que las personas que habitan en las otras regiones. La segunda región que más genera ingresos en promedio es el Sur. La que menos ingresos genera en promedio, es la Costa.

Existe una brecha salarial entre personas casadas y solteras. Se puede observar que los ingresos de las personas casadas son mayores a los ingresos de las personas solteras.