**Problem Set 2: Predicting Poverty**

1. **Introducción**

La sub-declaración de los ingresos por parte de los ciudadanos representa uno de los mayores retos del sistema fiscal, pues conocer el valor real de la renta individual es esencial para el cálculo los impuestos. De acuerdo con estudios realizados para Estados Unidos por el Servicio de Impuestos Internos (IRS), alrededor del 83,6 % de los impuestos se pagan de manera voluntaria y oportuna, con el reporte de ingresos más bajos que los reales como la principal causa de esta brecha. Ahora bien, resulta fundamental conocer el contexto colombiano, en donde el mercado laboral y el ingreso están permeados por variables sociales complejas. Así, a lo largo de este *Problem Set* se desarrolla un modelo de predicción de ingresos basado en características individuales usando datos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 del Departamento Administrativo Nacional de Estadística (DANE) para conocer información sociodemográfica de la población y señalar casos de fraude que podrían conducir a la reducción de la brecha en el país e identificar a las familias vulnerables que podrían necesitar mayor asistencia por parte del gobierno.

Las principales conclusiones indican que las variables de sexo y edad, tipo de ocupación, nivel de educación, experiencia y estrato socioeconómico, tienen impacto significativo en el salario de los incluidos estudiados y son clave para predecir sus ingresos por hora. Considerar estos resultados es crucial para la implementación de políticas fiscales efectivas y para mejorar la eficiencia del sistema de recaudo de impuestos.

**Nota:**

La base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: <https://github.com/Yilap/Repositorio_Taller1>

**Contexto**

En el mercado colombiano se presentan dos grandes fenómenos que afectan el recaudo de impuestos por parte de las entidades del Estado: i) la **evasión de impuestos** *(tax evasion)* que implica actos ilícitos por parte de los contribuyentes los cuales violan los deberes derivados de la relación jurídica tributaria - tales como presentar declaraciones verdaderas o mantener los libros comerciales regulares – y ii) la **elusión de impuestos** *(tax avoidance),* conocida como la práctica de actos mediante los cuales se influencian los canales de conexión para evitar la aplicación de ciertos gravámenes tributarios (Sentencia C-360 de 2016 Corte Constitucional de Colombia, 2016). La evasión de impuestos se genera en múltiples formas, entre las que se desatacan la omisión de ingresos, declaración de costos, deducciones y descuentos inexistentes, subvaloración de activos, mimetización ilegal de ingresos, entre otros. Estas acciones dificultan el ejercicio de las autoridades estatales para recaudar y hacer cumplir las obligaciones tributarias de los usuarios. Adicionalmente, se considera el fraude fiscal como delito, toda vez que el contribuyente tiene la intención de evadir el pago de sus obligaciones fiscales, como consecuencia de su comportamiento premeditado (Sentencia C-360 de 2016 Corte Constitucional de Colombia, 2016).

La evasión y elusión de impuestos ocasionan una pérdida para el gobierno de Colombia de entre COP $50 y $80 billones anuales. En 2021, según el Banco Interamericano de Desarrollo (BID), estos dos fenómenos representaron aproximadamente US$17 mil millones al año (o cerca de COP $68 mil millones) y, de acuerdo con Fedesarrollo, se pierden alrededor de 5,4 puntos porcentuales del PIB anual debido a estos factores. Sobre esto, es relevante mencionar que la mayor pérdida de recaudo se da por la evasión del impuesto de renta de las empresas, con un 3,4% del PIB; seguida de la evasión del IVA, que representa el 1,3% del PIB, y finalmente la evasión por concepto de impuesto de renta a personas, con cerca de un 0,7% del PIB (La República, 2022).

1. **Datos**
2. ***Descripción de las fuentes de datos***

Para el desarrollo de este Problem Set se utilizarán los datos de la Gran Encuesta Integrada de Hogares (GEIH) del Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta contiene información sobre las condiciones de empleo de las personas (si trabajan, en qué trabajan, cuánto ganan, fuentes de ingresos, si tienen seguridad social en salud o si están buscando empleo), adicional a las características generales de la población como sexo, edad, estado civil y nivel educativo (DANE, 2018). La GEIH consolida información no solo a nivel nacional sino a nivel regional, departamental, cabecera y ciudades capitales.

La GEIH es una fuente de información pertinente para el análisis de la evasión y elusión de impuestos en Colombia, ya que, al condensar información sociodemográfica sobre los individuos, los niveles de ingresos laborales y no laborales y la estructura de la fuerza de trabajo en el territorio nacional (tasas de ocupación, desempleo e informalidad laboral), permite identificar las características más relevantes que influyen sobre la renta personal. Usando estos datos como insumo, es posible predecir diferencias significativas entre los ingresos declarados y los reales.

1. ***Adquisición de los datos***

Para la obtención de los datos de la GEIH se utilizaron técnicas de web scraping. El conjunto de datos contiene todos los individuos muestreados en Bogotá y está disponible en el siguiente sitio web [https://ignaciomsarmiento.github.io/GEIH2018 sample/](https://ignaciomsarmiento.github.io/GEIH2018%20sample/). En este caso, la página web fuente de los datos divide la información en 10 “*chunks*” de datos. Para conocer la estructura de cada enlace fue necesario analizar el código HTML de la página web, logrando identificar su naturaleza dinámica. El hecho de ser dinámica plantea un reto para la extracción, pues el enlace original no es donde realmente reposa la información. Es necesario esperar a que cargue por completo para inspeccionarlo y encontrar enlace particular de la tabla a extraer, de lo contrario, no es posible detectarla mediante el código de R.

El web scraping se realizó usando el paquete *rvest*. En este caso, se identificó un patrón en los enlaces de cada data chunk, permitiendo importar la información por medio del loop a continuación y uniendo las 10 porciones de los datos en un solo data frame:

df\_list <- list()

for (i in 1:10) {

html\_i <- read\_html(paste0("https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_", i, ".html")) %>%

html\_table()

df\_i <- as.data.frame(html\_i)

df\_list[[i]] <- df\_i

}

GEIH <- do.call(rbind, df\_list)

1. ***Descripción del proceso de limpieza de datos***

La base de datos extraída contiene un total de 32.177 observaciones de todos los individuos muestreados en la GEIH en Bogotá y 178 variables. Para la limpieza de datos, se eliminan las variables que no resultan útiles para el modelo de predicción. Además, el análisis se centra únicamente en las personas empleadas mayores de 18 años, por lo que la muestra se limita a personas que cumplen estas características.

De acuerdo con la literatura económica, algunas de las variables más relevantes para la predicción de la renta individual son: edad, género, educación, experiencia, tipo de ocupación (relab) y estrato socioeconómico (estrato). A continuación, se justifica de manera detallada la inclusión de cada una de las variables en el modelo:

Cabe mencionar que previo a elegir las variables objeto de análisis se tuvo en cuenta la población en edad de trabajar (pet) toda vez que es fundamental contar con una segmentación por edades, ya que eso permite tener un panorama más claro para proceder con el análisis. La población en edad de trabajar representa aquellos individuos que pueden generar ingresos por concepto de trabajo y ser jefes de los hogares, haciendo que, esta variable sea necesaria para contar con un modelo objetivo y claro sobre cuál será la población para describir. Este segmento está constituido por las personas de 12 años y más en las zonas urbanas y 10 años y más en las zonas rurales. Además, se divide en población económicamente activa y población económicamente inactiva (DANE, s.f.).

* **Edad (Edad):** La edad de un individuo tiende a representar sus necesidades, oficios, intereses y preferencias. Por lo tanto, conocer la edad de los individuos nos permite generar un filtro para observar cuál es la población objetivo para cada investigación y planteamiento que se desee presentar. Por ejemplo, en este modelo de ingresos, los menores de edad no aportan información representativa, ya que cuentan usualmente con un jefe de hogar, quien percibe sus ingresos para manutención y demás necesidades. Por consiguiente, sus preferencias, oficios e intereses no serán analizadas en este espacio.
* **Género (Sex):** El género es fundamental en el análisis de los ingresos de los individuos, toda vez que, en el contexto colombiano, por ejemplo, existe una brecha entre hombres y mujeres en el momento de obtener trabajo y ganar un salario determinado, por lo tanto, ser hombre o mujer sí tiene influencia en la cantidad de ingresos que se perciben. Es por eso por lo que en el modelo tiene que estar presente esta variable, ya que ayudará a conocer el impacto en el salario dependiendo del género que tenga dicho individuo.

Adicionalmente, el enfoque de género tiene como objetivo identificar y caracterizar las particularidades contextuales y situaciones vividas por cualquier persona de acuerdo con su sexo, lo cual implica constructos sociales asociados, implicaciones y diferencias económicas, políticas, psicológicas, culturales y jurídicas, que pueden incidir en brechas sociales y eventuales situaciones de discriminación (DANE, 2022).

* **Educación (Educ):** La educación representa el nivel de cualificación de un individuo y, a su vez, muestra el nivel salarial que obtiene dado sus condiciones académicas. Por lo tanto, se asume que, entre más educación posea un individuo, es probable que sea es más competente y con ello, tiende a ser más productivo. Por lo tanto, contar con esta variable dentro del modelo permite analizar cuán importante es para saber el nivel de ingresos que puede obtener un individuo si aumenta uno o más años de educación.
* **Experiencia (Exp):** La experiencia permite conocer cuánto tiempo ha durado trabajando una persona, en este caso, la base de datos nos presenta los datos del tiempo que lleva trabajando la persona en la empresa actual. Esta variable es relevante para el análisis dado que aporta evidencia de la influencia que tiene la experiencia sobre el ingreso de una persona, es decir, entre más tiempo laboral posea, probablemente puede ser más productivo porque cuenta con más conocimientos, capacidades y habilidades para desempeñar sus actividades. Así pues, el modelo propuesto por Mincer (1974) sugiere que el salario de un individuo depende de su nivel educativo y su experiencia laboral, teniendo en cuenta otras variables relevantes que lo describan. Con esta ecuación, se pueden hacer ajustes para considerar las variaciones individuales. Por tanto, se puede afirmar que, los individuos con una mayor educación y experiencia tenderán a tener un mejor ingreso, lo que implica causalidad entre estas variables y el ingreso, lo que significa que entre mayores sean estas variables, mayores serán los ingresos.
* **Tipo de ocupación (Relab):** El tipo de ocupación se refiere a las categorías homogéneas de tareas que constituyen un conjunto de empleos, desempeñados por una persona, dadas sus capacidades y habilidades adquiridas por los años de educación y/o de experiencia adquirida, y por lo cual recibe un ingreso (DANE, 2005). Se puede inferir que, dependiendo del oficio, las personas obtendrán más o menos ingresos. Esta variable resume las principales tareas y deberes desempeñados en las ocupaciones y proporciona las categorías ocupacionales (DANE, 2015).
* **Estrato socioeconómico (Estrato):** El estrato socioeconómico es una clasificación en estratos de los inmuebles residenciales que deben recibir servicios públicos. Particularmente, es necesaria para cobrar de manera diferencial los servicios públicos domiciliarios, con el propósito de asignar subsidios y cobrar contribuciones. En este sentido, quienes tienen más capacidad económica pagan más por los servicios públicos y contribuyen para que los estratos bajos puedan pagar sus tarifas (DANE, s.f.).

Por otro lado, sirve para identificar geográficamente sectores con distintas características socioeconómicas para orientar la inversión pública, la asignación de programas sociales como mejoramiento de infraestructura de servicios públicos, vías, salud, saneamiento básico y servicios educativos y recreativos. Permite también el cobro de tarifas de impuesto predial diferenciales por estrato (DANE, s.f.). Lo anterior sirve como proxy para identificar en qué lugares se ubican los individuos con mayores o menores ingresos.

1. ***Análisis descriptivo de los datos (estadísticas descriptivas)***

Para iniciar con el análisis descriptivo de los datos, se procede a eliminar las observaciones de las personas menores de 18 años y las personas que se encuentran desocupadas (esto teniendo en cuenta el enunciado del Problem Set y que ya a los 18 años se puede laborar formalmente en Colombia). Adicionalmente, dado que en los puntos siguientes se utilizará la variable ingreso como logaritmo natural, se decidió eliminar las observaciones con valor de cero para que no existiesen datos incongruentes en la base. Se utilizaron los siguientes comandos:

GEIH <-GEIH[GEIH$age>=18,] y GEIH <- GEIH[GEIH$ocu == 1, ].

Posteriormente, se renombra la variable de máximo nivel de educación para tener mayor claridad. Este proceso se hace con el comando GEIH <- rename(GEIH, educ = p6210). En seguida, las variables, educación y tipo de ocupación, se establecen como categóricas de la siguiente manera:

GEIH$educ <- factor(GEIH$educ)

class(GEIH$educ)

GEIH$relab <- factor(GEIH$relab)

class(GEIH$relab)# Cálculo de la experiencia potencial

Por otro lado, en primer lugar, se estiman los años de educación dependiendo del máximo nivel alcanzado como se observa a continuación:

GEIH$añoseduc <- ifelse(GEIH$educ == 3, 5,

ifelse(GEIH$educ == 4, 9,

ifelse(GEIH$educ == 5, 11,

ifelse(GEIH$educ == 6, 16,

ifelse(GEIH$educ == 9, 0, 0)))))

Se aplica la fórmula de experiencia potencial, en la cual los valores negativos se aproximan a 0 experiencia y se eliminan las personas que tengan una experiencia de 0 años.

GEIH$experp <- GEIH$age - 5 - GEIH$añoseduc

GEIH$experp <- ifelse(GEIH$experp < 0, 0, GEIH$experp)

GEIH<- GEIH[GEIH$experp>0,]

Ahora bien, para el cálculo de las horas totales trabajadas, se suman las horas trabadas en el empleo principal y secundario en una sola variable denominada **horast,** aplicando los siguientes comandos:

GEIH$hoursWorkActualSecondJob <- ifelse(is.na(GEIH$hoursWorkActualSecondJob), 0, GEIH$hoursWorkActualSecondJob)

GEIH$horast <- GEIH$hoursWorkUsual + GEIH$hoursWorkActualSecondJob

Seguido a esto, se eliminan las personas que tengan un ingreso total de 0.

GEIH<- GEIH[GEIH$ingtot>0,]

Y para obtener el salario por hora se realiza la siguiente operación: (Ingresos mensuales, por 12 meses, dividido en las horas semanales trabajadas por 52 semanas del año).

GEIH$inghora <- (GEIH$ingtot\*12)/(GEIH$horast\*52)

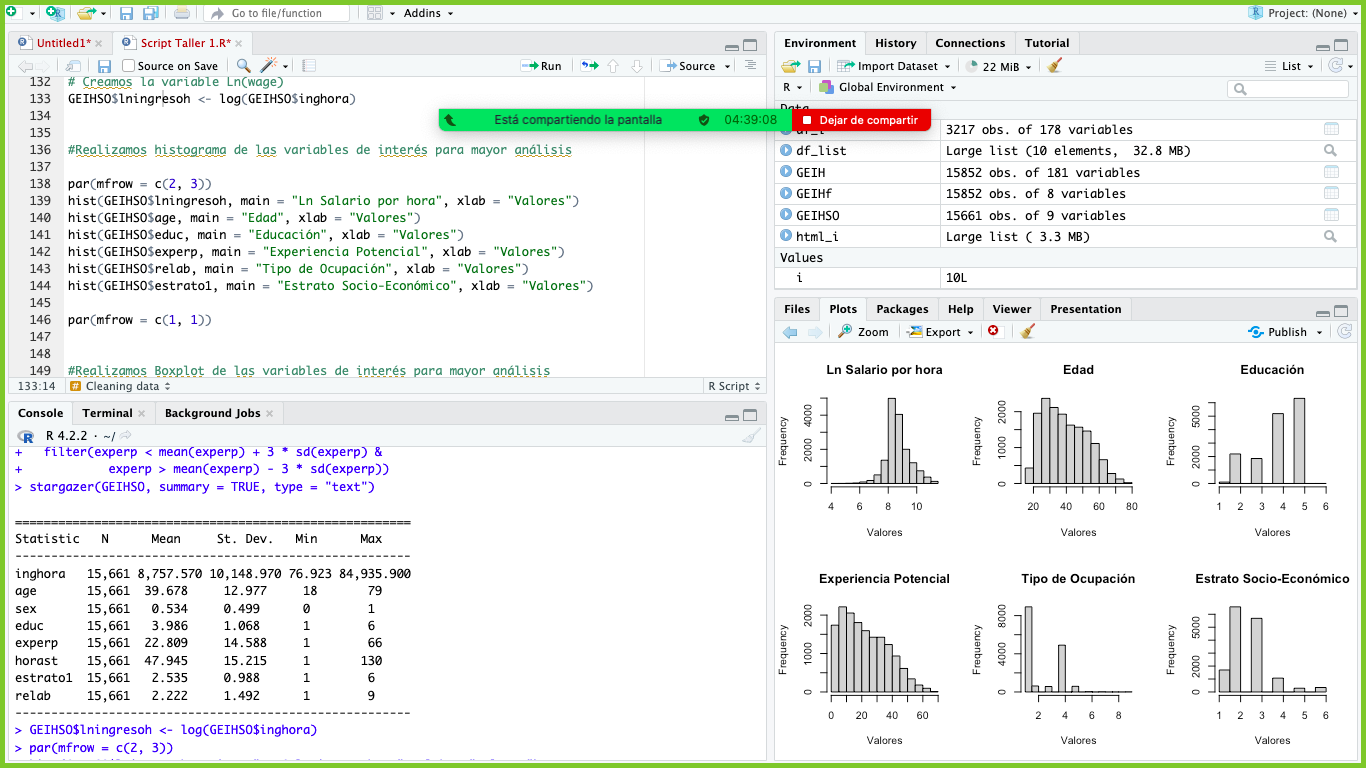
Finalmente, se hace un subset de las variables de interés objeto de análisis para obtener subconjuntos, como se evidencia a continuación:

GEIHf <-subset(GEIH, select = c("inghora","age","sex","educ","experp","horast","estrato1","relab"))

GEIHf <- na.omit(GEIHf)

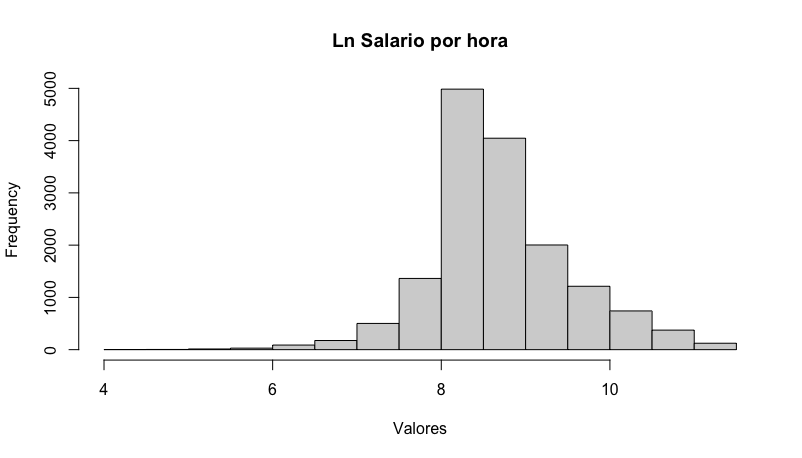
Y se identificaron y eliminaron los outliers, estos serán los que estén 3 desviaciones estándar alejados de la media, con lo que se creó la nueva base de datos denominada GEIHSO.

Con lo anterior, se procede a realizar las respectivas estadísticas descriptivas de nuestras variables de interés ("inghora","age","sex","educ","experp","horast","estrato1","relab")), con lo que se puede inferir que se tienen 15.661 observaciones con 8 variables.



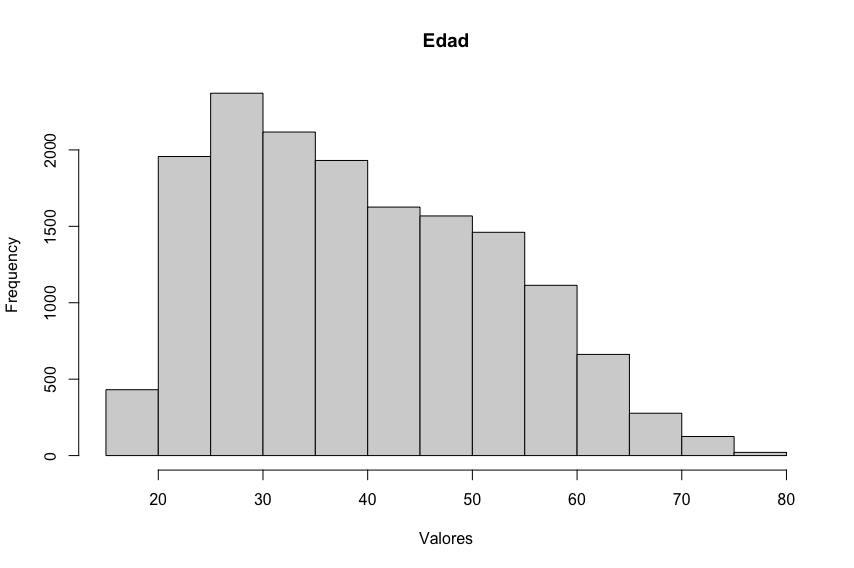
* **Ingreso por hora**

La Ingreso por hora se refiere al salario devengado en promedio por las personas. Se puede observar una asimetría hacia la derecha. El salario mínimo es 76 pesos y el máximo es de 84.935 pesos por hora, con un promedio de 8.757 pesos.

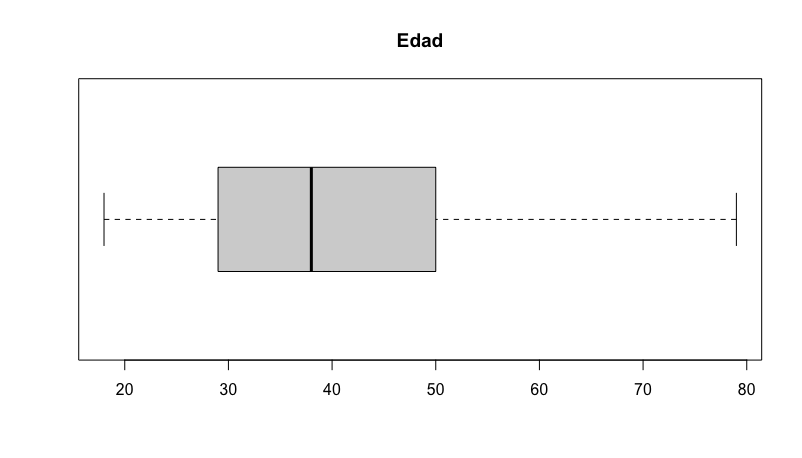
****

* **Edad**

La Edad hace referencia a que la variable maneja números enteros. Se puede observar una asimetría hacia la izquierda y gran parte de los individuos están entre los 20 y 40 años. Adicionalmente, se puede inferir que, de acuerdo con la limpieza de la base de datos y los filtros realizados, la edad mínima es de 18 años, la edad máxima es de 79 años y el promedio de edad 39 años.



En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre 30 y 50 años con una media situada en 39 años.



* **Sexo**

La variable es categórica, toda vez que hace referencia al género de los individuos: Hombre o Mujer. Por lo tanto, cuenta con dos niveles: 1= Hombre y 0 = Mujer, evidenciando que, del total de la muestra (15.661), el 53,4% es equivalente a los hombres, es decir, 8.363; el restante 46,6% equivale a las mujeres, es decir, 7.298. Lo anterior, evidencia que los hombres tuvieron mayor participación en comparación con las mujeres.

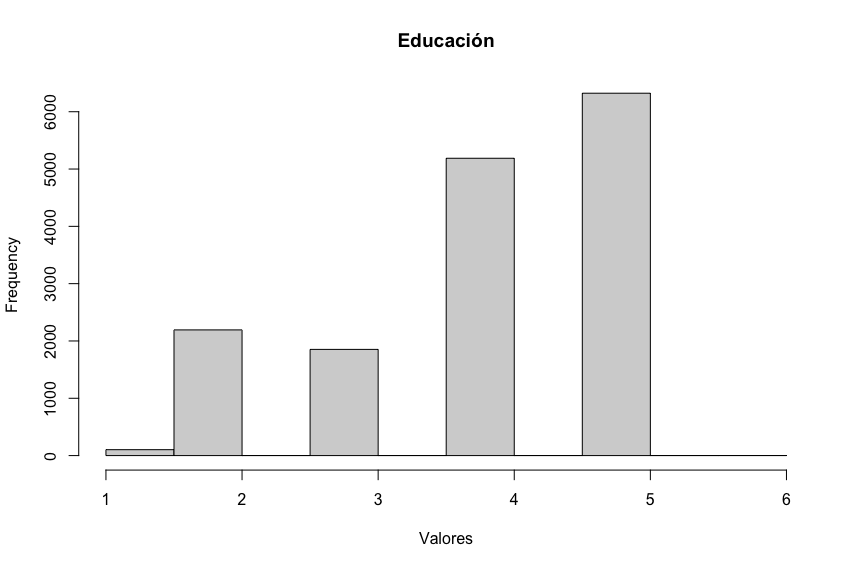
* **Educación**

Esta variable responde a la pregunta ¿Cuál es el nivel educativo más alto alcanzado por .... y el último año o grado aprobado en este nivel? Con esta pregunta se trata de obtener el nivel educativo más alto alcanzado y el último grado de ese nivel. Si bien nos indica que la variable cuenta con número enteros, se pude conocer que dichos números representan diferentes variables, tales como:

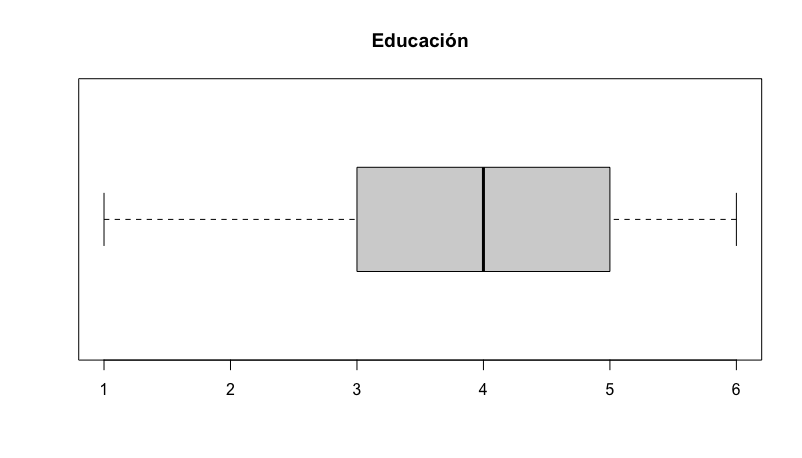
1. Ninguno
2. Preescolar
3. Básica primaria (1º a 5º)
4. Básica secundaria (6º a 9º)
5. Media (10º a 13º)
6. Superior o universitaria

9. No sabe, no informa

Por lo tanto, podemos deducir que el promedio de los encuestados cuenta con educación media. Sin embargo, la categoría con mayor frecuencia es la 6, es decir, superior o universitaria, tal como se demuestra en el histograma. Podemos inferir que gran parte de la muestra cuenta con una educación superior o universitaria.

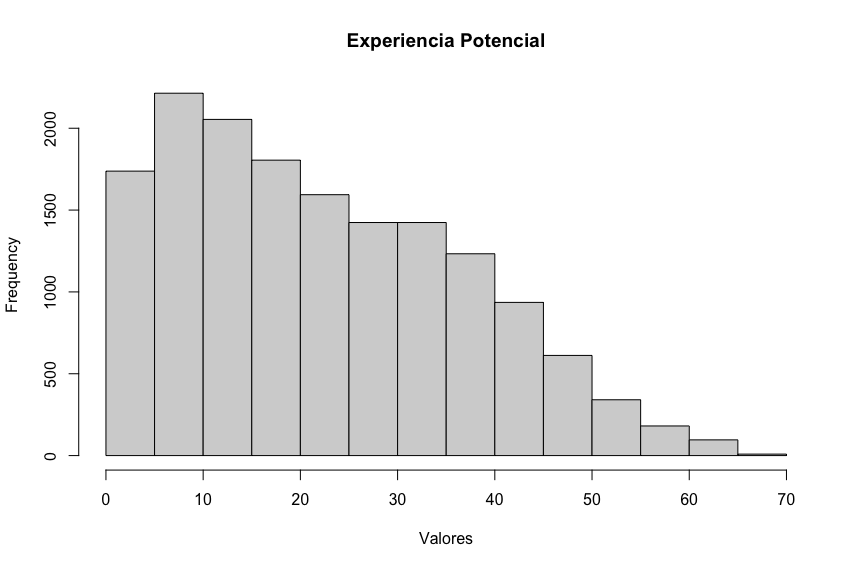
****

En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre 4 y 6 con una media situada en 5.

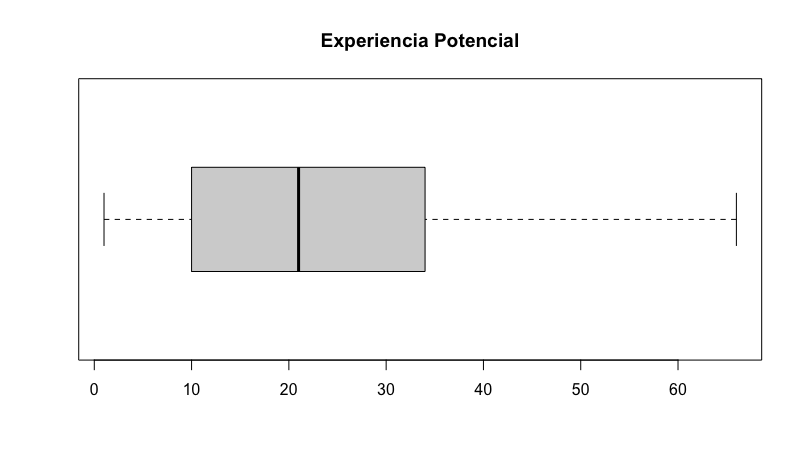
****

* **Experiencia**

Esta variable es numérica, en donde se mide el número de años que el individuo lleva en su último trabajo. Por lo tanto, se puede observar que la frecuencia de los datos es asimétrica hacia la izquierda, el dato mínimo es 1, es decir, un año, el máximo son 66 años, por consiguiente, el dato más frecuente es que los individuos lleven un año o menos en su oficio actual.



En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre 10 y 35 años con una media situada en 22 años.

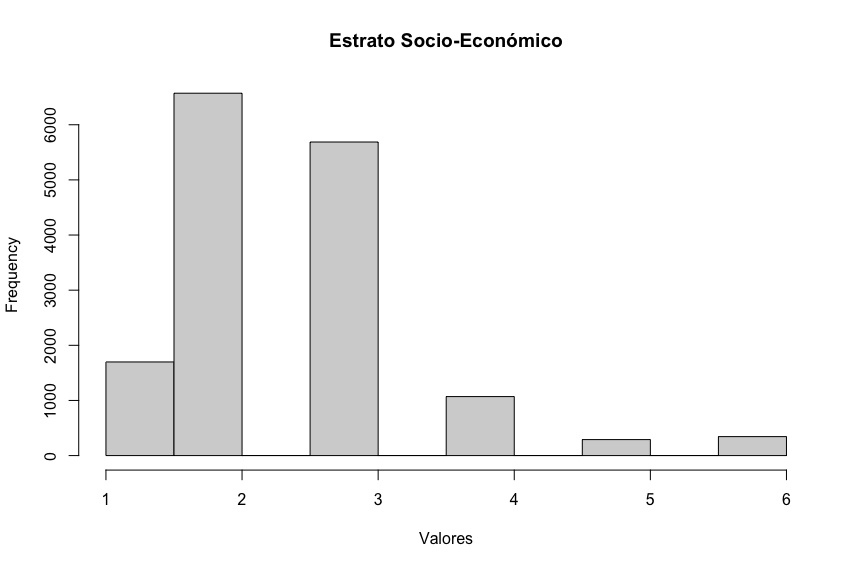


* **Horas totales**

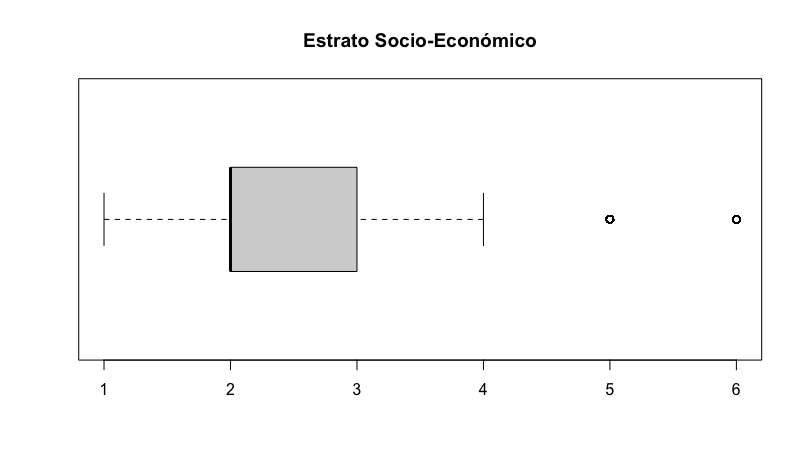
La variable horas totales es cuantitativa que mide el total de las horas trabajadas en promedio por una persona. En razón a esto, se puede observar que las horas totales mínimas laboradas es de 1 y las máximas son de 130 horas. El promedio es de 47 horas.

* **Estrato socioeconómico**

El estrato se refiere a una variable que toma valores entre 1 y 6, donde 1 es el estrato más bajo y 6 el estrato más alto. El promedio de los datos se encuentra en el estrato 2, seguido del estrato 3, luego estrato 1, y finalmente, los estratos 4, 5 y 6.



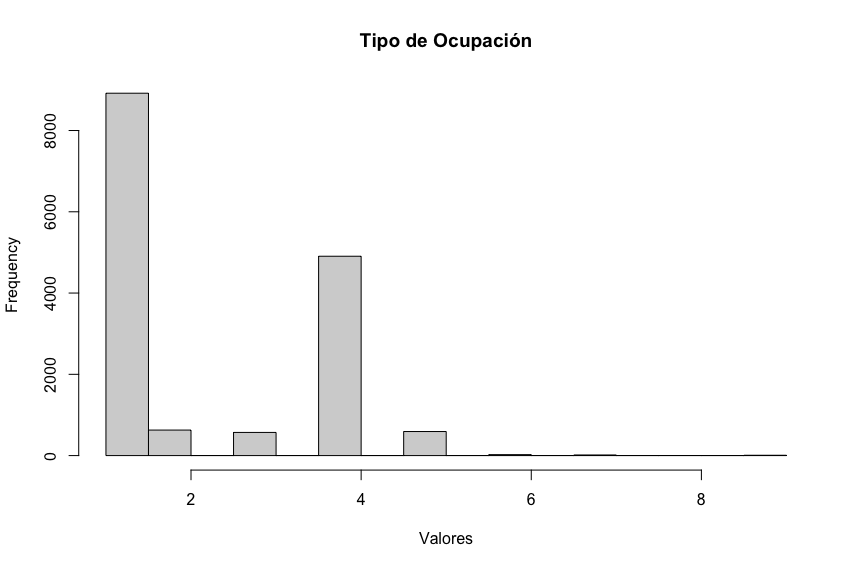
En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre los estratos 2 y 3 con una media situada en el estrato 2. Es importante mencionar que, se presentan outliers, representados en los estratos 5 y 6, por ser los estratos donde se concentran menos observaciones.



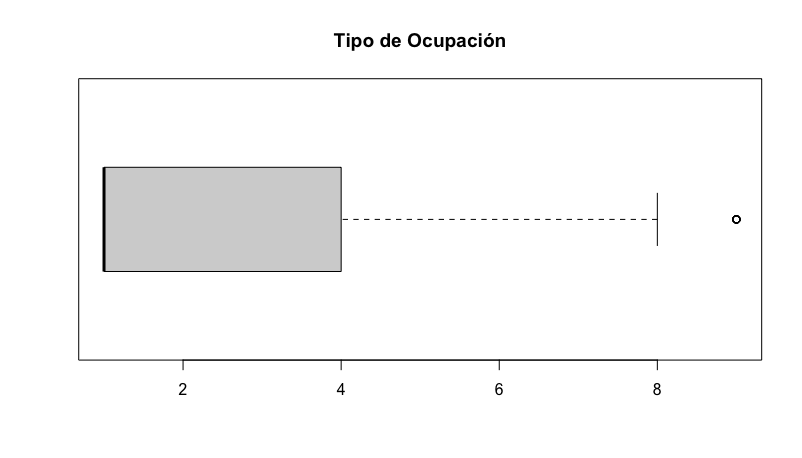
* **Tipo de ocupación**

Como se puede observar, el Tipo de Ocupación hace referencia a una variable categórica, en donde 1 es Obrero o empleado de empresa particular; 2 Obrero o empleado del gobierno; 3 Empleado doméstico; 4 Trabajador por cuenta propia; 5 Patrón o empleador; 6 Trabajador familiar sin remuneración; 7 Trabajador sin remuneración en empresas o negocios de otros hogares; 8 Jornalero o peón; y 9 Otro.

En este sentido, se observa que, la mayor parte de la población trabaja como obrero o empleado de empresa particular o como trabajadores por cuenta propia.



En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre los tipos de ocupación 1 y 4, con una media situada en el tipo de ocupación 1. Es importante mencionar que, se presentan outliers, representados en el tipo de ocupación 9, es decir, en otros.



1. **Modelos y resultados**
2. ***Clasificación de modelos***
3. ***Modelo de regresión de los ingresos***
4. **Modelo Final**
5. **Conclusiones y recomendaciones**
6. **Bibliografía**

Alcaldía de Bogotá., (2016). Documentos para DELITOS CONTRA EL ORDEN ECONÓMICO Y SOCIAL: Evasión Fiscal. Recuperado de <https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=32520&cadena=>