**Taller 1.**

**Predicción de Ingresos.**

**Integrantes:** Isabella Mendez Pedraza. Cód.:

Manuela Ojeda Ojeda. Cód.:

Juan Sebastian Tellez Melo. Cód.: 201513710

Andres Mauricio Palacio Lugo. Cód.: 201618843

1. **Introducción.**

Para determinar el valor de los impuestos que debe pagar cada persona es fundamental la exactitud en la declaración de los ingresos.        Sin embargo, el fraude fiscal de todos los tipos siempre ha sido un problema que está muy presente. Menos del 90% de los impuestos son pagados voluntariamente y en el tiempo estimado en USA. La razón de esta brecha es que muchas personas no reportan correctamente sus ingresos.

Un modelo predictivo del ingreso podría ayudar a señalar casos de fraude para reducir la brecha. Además, podría ayudar a identificar personas en condiciones de vulnerabilidad que puedan necesitar un apoyo o ayuda adicional.

1. **Datos.**

La Gran Encuesta Integrada de Hogares (GEIH) del DANE es una encuesta mediante la que, además de las características generales de la población (como la edad, sexo, nivel educativo, entre otros), se obtiene información acerca de las condiciones de empleo de las personas (si están empleados, cantidad de horas que trabajan, si tienen seguridad social) y su nivel y fuentes de ingresos. Con esta información se generan indicadores del mercado laboral en Colombia como la tasa de ocupación, la rama laboral en la que se desempeñan los colombianos y la remuneración, y el comportamiento del mercado laboral para grupos poblacionales específicos como los jóvenes.

**Scrapping**

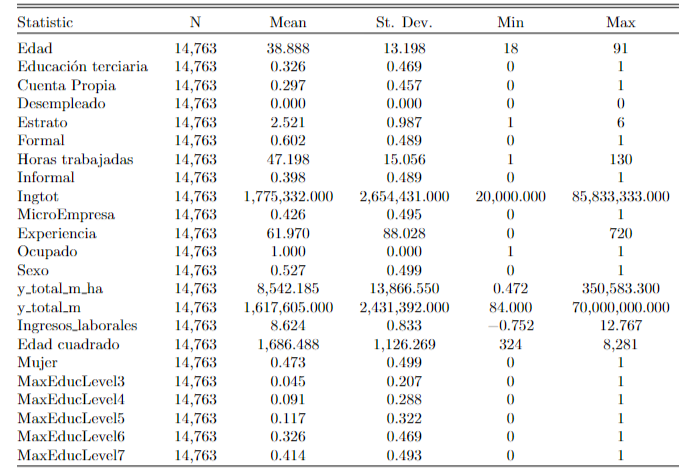
Para el análisis del presente documento, se usaron datos extraídos de la GEIH con una serie de modificaciones hechas por profesores de la Facultad de Economía de la Universidad de Los Andes. Estos datos fueron publicados en 10 páginas de internet diferentes por lo que fue necesario acceder a ellos mediante métodos de *web scraping*, utilizando el software estadístico de R y el código utilizado para ello se encuentra en el repositorio descrito al inicio del documento en el script denominado “*2\_Data\_cleaning*”. En este procesamiento de los datos desde las páginas de internet se identificó que los datos, aunque no tenían restricciones de seguridad, si se encontraban almacenados en objetos que requirieron revisar en detalle los recursos de red de cada página de internet. Una vez se logró obtener los datos, se creó un *loop* en R que permitió descargarlos y almacenarlos unidos en tables locales.

**Limpieza**

En un primer acercamiento con los datos optamos por tomar las variables que consideramos relevantes para el análisis, aquellas que pudieran explicar o relacionarse con el nivel de ingresos de las personas y también aquellas variables características de las personas o del contexto de la muestra. Iniciamos la limpieza eliminando las personas que son desempleadas o menores de 18 años. Para tener mayor claridad sobre los datos a manejar y depurar la base, calculamos el porcentaje de missing values por variable y eliminamos aquellas que tuvieran un porcentaje mayor al 50%. Luego, eliminamos los NA’s restantes. Finalmente, para dar inicio al análisis, generamos las variables que necesitábamos, como “Ingresos\_laborales” y definimos las varaibles catégoricas.

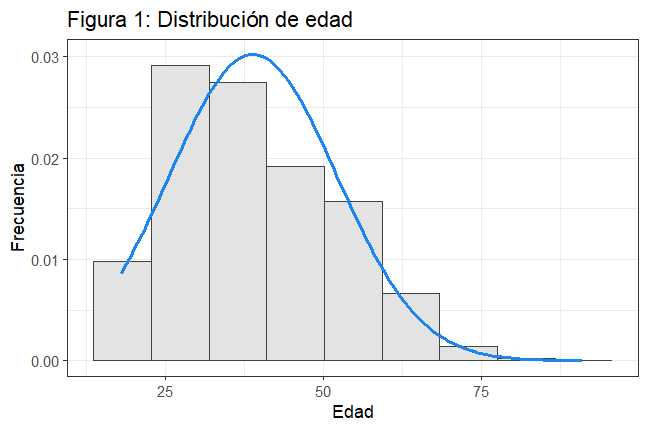
**Estadístmicas descriptivas**

Tabla 1. Estadísticas descriptivas

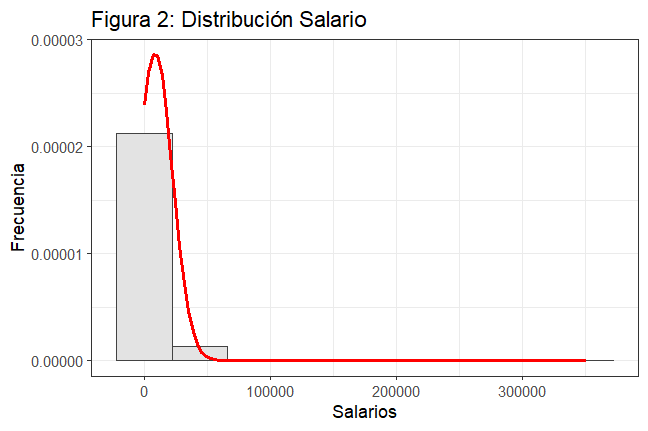
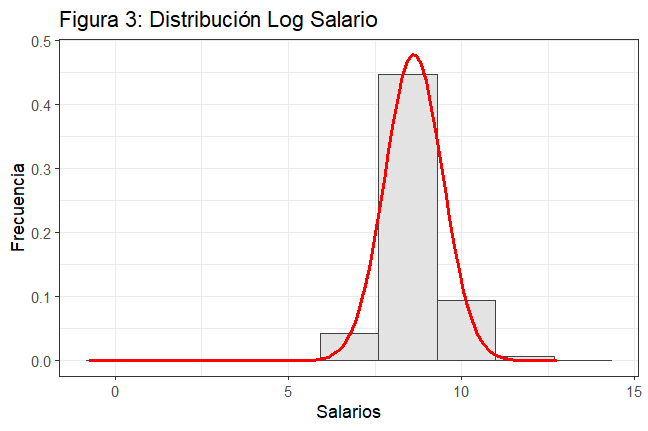


Como parte de la limpieza de datos, se restringió la información a sujetos con edades a partir de los 18 años y que estuviesen ocupados (utilizando la variable “ocu =1”). Así, llegamos a una muestra de 14,763 observaciones. De esta muestra, el 47.3% corresponde a mujeres mientras que el 52.7% son hombres.

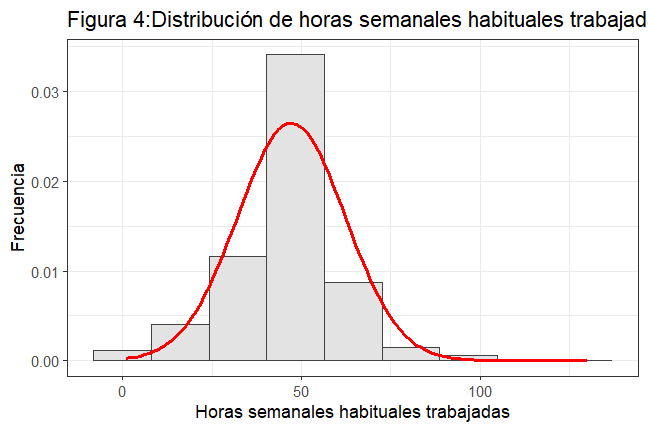
La edad promedio de las personas es 39 años con una desviación estándar de 13 años y la persona ocupada con mayor edad tiene 91 años.



En la figura 1 observamos la distribución por edad, allí se evidencia que solo tenemos personas mayores de edad y hay muy pocas personas que tienen más de 75 años. También se observa que las edades con mayor frecuencia están entre los 20 años y 30 años.



En la figura 2 se observa la distribución de los salarios sin ningún tratamiento, sin embargo, en la figura 3 se evidencia cómo es la distribución del logaritmo del salario de los individuos, que es la variable que se utilizará en las estimaciones de este documento. Se observa que la media del logaritmo del salario está entre $8 y $9 mil pesos por hora.



A partir de la tabla 1 y la figura 4 se observa que en promedio los individuos trabajan 47 horas habitualmente en la semana, la persona que menos trabaja es 1 hora y la que más hora trabaja llega a 130 horas. Los individuos trabajan con mayor frecuencia entre 40 y 50 horas habitualmente en la semana.

1. **Perfil Edad-Salario**
2. **Brecha salarial por género**
3. **Predicción de ingresos**