Problem Set 1:

Predicting Income

**Integrantes:** \_\_\_\_\_\_\_\_\_\_

## **Introducción**

Ta

## **Datos**

Los datos de la encuesta “Medición de Pobreza Monetaria y Desigualdad 2018” es elaborada a partir de la “Gran Encuesta Integrada de Hogares” por el DANE. Usando los datos de esta encuesta se pretende predecir el nivel de ingreso laboral de ocupados mayores de 18 años radicados en la ciudad de Bogotá, con el fin de desarrollar un modelo predictivo que pueda generar alertas relevantes para las autoridades encargadas de la recaudación de impuestos, utilizando ciertos parámetros específicos.

Para recopilar los datos, realizamos web scraping seleccionando 10 archivos de la web que contienen información de la encuesta para cada mes de 2018 en la ciudad de Bogotá. Es importante destacar que la página de la cual se extrajo los datos no impone restricciones para el uso de web scraping. Inicialmente, se cuentan con 32.177 observaciones para Bogotá, pero nos enfocamos en 24.568 personas mayores de edad y \_\_\_\_ de la población ocupada. Luego, eliminamos las observaciones que tenían valores faltantes en la variable de salario nominal mensual, lo que resultó en la eliminación de \_\_\_\_ observaciones debido al alto porcentaje de datos faltantes.

Para tratar los valores atípicos en los salarios, aplicamos una técnica llamada "Winsorización". Esta técnica implica mantener las observaciones atípicas y ajustar sus valores de acuerdo con los percentiles elegidos. En este caso, las observaciones atípicas se ajustaron de manera que el 1% más bajo de la muestra tomara el valor del percentil 1, y el 1% más alto de la distribución de los salarios tomara el valor del percentil 99. Como resultado de este proceso, nuestra muestra de estudio final contiene 9,892 observaciones.