**Problem set 1**

**1.a.i.**

**ii.** La base de datos obtenida se trata de la encuesta GEIH del 2018 para la ciudad de Bogotá, esta es la gran encueta integrada de hogares y contiene información sobre las condiciones de empleo de las personas, además de sus fuentes de ingreso y características generales de la población. En cuanto al acceder a los datos y hacer el scrapping, hubo inconvenientes en la medida que la página ofrecida por el profesor no cargaba en Chrome, por lo que se utilizo el buscador Mozilla Firefox para acceder al url escondido en la sección de red (network) al momento de inspeccionar los chunks.

**b.**

Estadísticas descriptivas:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| directorio | ingtot | sex | age | clase | college | cuentaPropia | depto | formal | maxEducLevel | oficio | relab | totalHoursWorked |
| Min. : 1 | Min. : 0 | Min. :1.00 | Min. :18.00 | Min. :1 | Min. :1.000 | Min. :1.000 | Min. :1 | Min. :1.000 | Min. :1.000 | Min. : 1.00 | Min. :1.000 | Min. : 1.0 |
| 1st Qu.:2338 | 1st Qu.: 800000 | 1st Qu.:1.00 | 1st Qu.:28.00 | 1st Qu.:1 | 1st Qu.:1.000 | 1st Qu.:1.000 | 1st Qu.:1 | 1st Qu.:1.000 | 1st Qu.:4.000 | 1st Qu.:24.00 | 1st Qu.:1.000 | 1st Qu.: 40.0 |
| Median :4636 | Median : 1051160 | Median :2.00 | Median :38.00 | Median :1 | Median :1.000 | Median :1.000 | Median :1 | Median :2.000 | Median :5.000 | Median :36.00 | Median :1.000 | Median : 48.0 |
| Mean :4624 | Mean : 1769379 | Mean :1.53 | Mean :39.44 | Mean :1 | Mean :1.319 | Mean :1.309 | Mean :1 | Mean :1.587 | Mean :4.954 | Mean :39.45 | Mean :2.268 | Mean : 47.4 |
| 3rd Qu.:6933 | 3rd Qu.: 1723158 | 3rd Qu.:2.00 | 3rd Qu.:50.00 | 3rd Qu.:1 | 3rd Qu.:2.000 | 3rd Qu.:2.000 | 3rd Qu.:1 | 3rd Qu.:2.000 | 3rd Qu.:6.000 | 3rd Qu.:52.00 | 3rd Qu.:4.000 | 3rd Qu.: 50.0 |
| Max. :9214 | Max. :85833333 | Max. :2.00 | Max. :94.00 | Max. :1 | Max. :2.000 | Max. :2.000 | Max. :1 | Max. :2.000 | Max. :6.000 | Max. :80.00 | Max. :9.000 | Max. :130.0 |
| NA:  0 | NA:  0 | NA:  0 | NA:  0 | NA:0 | NA:  0 | NA:  0 | NA:0 | NA:  0 | NA's :1 | NA:  0 | NA:0 | NA:  0 |

A partir de la tabla anterior se observan las principales características de las variables que se deciden utilizar para realizar el problema set. En primer lugar, la variable de ingtot hace referencia al ingreso total de la persona encuestada, se escogió esta como la medida apropiada del ingreso ya que recoge las diferentes formas y tipos de ingreso que una persona recibe, dando así una visión general de este rubro.

Continuando, esta variable nos indica que en promedio los bogotanos tienen un ingreso total de $1´769.379 de pesos colombianos teniendo como valor mínimo el 0 (se entiende por construcción de la base de datos que se presentan situaciones donde las personas son ocupados pero que no perciben ingresos por este trabajo) y como valor máximo $85´833.333. Continuando, nos encontramos con que las personas dentro de la muestra tienen en promedio 39.4 años de edad y que hay ligeramente más hombres encuestados que mujeres (media de 1.53, siendo el valor de 2 igual a hombre y de 1 igual a mujer). Junto a esto, se observa que en promedio las personas trabajaron 47.4 horas la semana anterior a ser encuestadas, lo cual se toma como el promedio trabajado cada semana.

Además, se presentan variables que se considera pueden ser buenos predictores del ingreso como lo son: clase, la cual toma el valor de 1 si la persona encuestada reside en zona urbana y 0 si reside en zona rural (en las observaciones que se tienen solo hay residentes de la zona urbana); collegue que indica si la persona tiene educación terciaria (=1); cuentaPropia, que indica si la persona es autoempleada (=1) o no, formal, que toma el valor de 1 si la persona trabaja dentro del sector formal de la economía; maxEducLevel, que es el nivel de educación máximo que obtiene la persona; y oficio y relab que son, respectivamente, la ocupación y el tipo de ocupación de la persona.

Según las gráficas presentes en la carpeta views, se observa que la distribución del ingreso total dentro de la muestra esta muy concentrada en la zona izquierda del plano, lo cual muestra que muchas personas ingresan montos mucho menores en relación con quienes expanden el histograma a la derecha, esto junto a los valores mínimo y máximo soportan la existencia de una fuerte inequidad en el ingreso. Respecto a los diagramas de dispersión, se observa como las horas trabajadas se correlacionan de forma positiva con el ingreso total, pero esta relación no es muy fuerte por lo que se puede decir que hasta cierto punto de las horas trabajas dan mayores ingresos. En cuanto a la relación de ingreso total y edad, estos también se relacionan positivamente, pero hay una tendencia más clara a que en los 50 años aproximadamente se encuentran las personas con mayores niveles de ingreso total. En cuanto al histograma de nivel de educación máximo e ingreso total, se puede observar la clara tendencia de que a mayor sea el nivel de educación alcanzado, mayor es el ingreso total que se percibe. Por último, se debe mencionar que en los 3 gráficos de dispersión ya presentados se diferencian a los hombres de las mujeres por el color de los puntos, los hombres toman el color azul claro y las mujeres el azul oscuro, y con esta diferenciación es posible ver como los hombres tienden a tener mayores ingresos que las mujeres, así estas tengan el mismo nivel de educación, trabajen las mismas horas o tengan la misma edad.