Entrega M4 y M5 DCSCHD-UNSAM

Gonzalo Azuaga

# Carga de paquetes

pacman::p\_load(tidyverse, caret, DataExplorer, skimr, flextable, gridExtra, grid, ppcor, glmnet)

# Importación de base de datos

datos <- read\_csv("bases/01\_2020\_II\_eph\_ocupados.csv")

head(datos, n = 10) # visualización de los primeros 10 casos

## # A tibble: 10 x 14  
## CODUSU NRO\_HOGAR NIVEL\_ED CH03 CH04 CH06 CH07 PP04A CAT\_OCUP INTENSI   
## <chr> <dbl> <chr> <chr> <chr> <dbl> <chr> <chr> <chr> <chr>   
## 1 TQRMNO~ 1 Superior ~ Jefe/a Varon 38 unid~ ...e~ Obrero ~ Ocupado~  
## 2 TQRMNO~ 1 Superior ~ Conyu~ Mujer 35 unid~ ...e~ Obrero ~ Sobreoc~  
## 3 TQRMNO~ 1 Secundari~ Jefe/a Varon 56 solt~ ...p~ Cuenta ~ Ocupado~  
## 4 TQRMNO~ 1 Secundari~ Hijo/~ Varon 43 sepa~ ...e~ Obrero ~ Ocupado~  
## 5 TQRMNO~ 1 Superior ~ Jefe/a Varon 58 unid~ ...p~ Cuenta ~ Sobreoc~  
## 6 TQRMNO~ 1 Superior ~ Conyu~ Mujer 49 unid~ ...e~ Obrero ~ Ocupado~  
## 7 TQRMNO~ 1 Superior ~ Jefe/a Mujer 50 solt~ ...p~ Cuenta ~ Ocupado~  
## 8 TQRMNO~ 1 Primaria ~ Jefe/a Mujer 64 solt~ ...p~ Obrero ~ Ocupado~  
## 9 TQRMNO~ 1 Superior ~ Hijo/~ Varon 28 solt~ ...p~ Cuenta ~ Ocupado~  
## 10 TQRMNO~ 1 Secundari~ Hijo/~ Mujer 38 solt~ ...p~ Obrero ~ Ocupado~  
## # ... with 4 more variables: PP3E\_TOT <dbl>, CATEGORIA <chr>,  
## # CALIFICACION <chr>, P21 <dbl>

El dataset a utilizar es un subconjunto de datos de la Encuesta Permanente de Hogares correspondiente (EPH) al II trimestre del 2020. Sólo están registradas las personas que declararon haber trabajado al menos una hora durante la semana anterior al operativo, es decir *ocupadas*, y se han seleccionado algunas variables para el análisis. Dicho dataset cuenta con 12803 filas, cada una de ellas representa a un individuo, y 14 variables. A continuación, se renombran las variables y se presenta una tabla con las definiciones de cada una de ellas.

# Renombramiento de variables

datos <- datos %>%   
 rename(Nivel\_ed = NIVEL\_ED,  
 Nro\_hogar = NRO\_HOGAR,  
 Codusu = CODUSU,  
 Relacion = CH03,  
 Sexo = CH04,  
 Edad = CH06,  
 Estado = CH07,  
 Tipo = PP04A,  
 Intensidad = INTENSI,  
 Categoria\_ocup = CAT\_OCUP,  
 Total\_horas = PP3E\_TOT,  
 Categoria = CATEGORIA,  
 Calificacion = CALIFICACION,  
 Ingreso = P21)

data <- data.frame(Variable = names(datos)) %>%   
 mutate(Descripción = case\_when(  
 Variable == 'Nivel\_ed' ~ 'máximo nivel educativo alcanzado',  
 Variable == 'Nro\_hogar' ~ 'refiere al hogar dentro de la vivienda. Puede haber más de un hogar en una vivienda',  
 Variable == 'Codusu' ~ 'número de identificación que permite relacionar una vivienda con hogares y personas',  
 Variable == 'Relacion' ~ 'relación de parentesco con el jefe/a de hogar',  
 Variable == 'Estado' ~ 'estado conyugal',  
 Variable == 'Tipo' ~ 'sector del establecimiento donde trabaja (público, privado)',  
 Variable == 'Intensidad' ~ 'intensidad de la ocupación (ocupado pleno, sobreocupado, subocupado)',  
 Variable == 'Categoria\_ocup' ~'categoría ocupacional',  
 Variable == 'Total\_horas' ~ 'cantidad de horas que trabajó la semana anterior en la ocupación principal',  
 Variable == 'Categoria' ~ 'carácter de la ocupación principal (basado en el CNO)',  
 Variable == 'Calificacion' ~ 'calificación de la ocupación principal (basado en el CNO)',  
 Variable == 'Ingreso' ~ 'Ingreso de la ocupación principal',  
 Variable == 'Sexo' ~'sexo',  
 Variable == 'Edad' ~'edad en años cumplidos'))   
  
#Cambio de formato de la tabla compatible con Word  
flextable((data),   
 col\_keys = c("Variable", "Descripción" )) %>%   
 add\_header\_lines(.,values = c("Tabla 1. Definición de variables")) %>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., width =.5, layout = "autofit")

| Tabla 1. Definición de variables | |
| --- | --- |
| Variable | Descripción |
| Codusu | número de identificación que permite relacionar una vivienda con hogares y personas |
| Nro\_hogar | refiere al hogar dentro de la vivienda. Puede haber más de un hogar en una vivienda |
| Nivel\_ed | máximo nivel educativo alcanzado |
| Relacion | relación de parentesco con el jefe/a de hogar |
| Sexo | sexo |
| Edad | edad en años cumplidos |
| Estado | estado conyugal |
| Tipo | sector del establecimiento donde trabaja (público, privado) |
| Categoria\_ocup | categoría ocupacional |
| Intensidad | intensidad de la ocupación (ocupado pleno, sobreocupado, subocupado) |
| Total\_horas | cantidad de horas que trabajó la semana anterior en la ocupación principal |
| Categoria | carácter de la ocupación principal (basado en el CNO) |
| Calificacion | calificación de la ocupación principal (basado en el CNO) |
| Ingreso | Ingreso de la ocupación principal |
| Fuente: EPH 2º Trimestre 2020 | |

# Análisis exploratorio

df\_skim\_ch <- skim(datos) %>%   
 filter(skim\_type == "character") %>%   
 rename(Variable = skim\_variable,  
 Tipo = skim\_type,  
 Vacíos = character.empty,  
 N\_únicos = character.n\_unique,  
 Espacios\_vacíos = character.whitespace,  
 NA\_s = n\_missing) %>%   
 dplyr::select(Tipo, Variable,N\_únicos, Espacios\_vacíos, NA\_s)  
  
  
flextable((df\_skim\_ch),   
 col\_keys = names(df\_skim\_ch)) %>%   
 add\_header\_lines(.,values = c("Cuadro 2. Análisis exploratorio de variables categóricas"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 2. Análisis exploratorio de variables categóricas | | | | |
| --- | --- | --- | --- | --- |
| Tipo | Variable | N\_únicos | Espacios\_vacíos | NA\_s |
| character | Codusu | 8,338 | 0 | 0 |
| character | Nivel\_ed | 7 | 0 | 0 |
| character | Relacion | 10 | 0 | 0 |
| character | Sexo | 2 | 0 | 0 |
| character | Estado | 5 | 0 | 0 |
| character | Tipo | 4 | 0 | 0 |
| character | Categoria\_ocup | 4 | 0 | 0 |
| character | Intensidad | 4 | 0 | 0 |
| character | Categoria | 51 | 0 | 0 |
| character | Calificacion | 7 | 0 | 0 |
| Fuente: EPH 2º Trimestre 2020 | | | | |

df\_skim\_num <- skim(datos) %>%   
 filter(skim\_type == "numeric") %>%   
 mutate(across(where(is.numeric),round, 2)) %>%   
 rename(Variable = skim\_variable,  
 Tipo = skim\_type,  
 Media = numeric.mean,   
 Desvío\_std=numeric.sd,   
 Mín=numeric.p0,  
 Q1=numeric.p25,  
 Mediana=numeric.p50,  
 Q3=numeric.p75,  
 Máx=numeric.p100,   
 NA\_s =n\_missing) %>%   
 dplyr::select(Tipo, Variable, !starts\_with("character")) %>%   
 dplyr::select(-complete\_rate)  
  
flextable((df\_skim\_num),   
 col\_keys = names(df\_skim\_num)) %>%   
 add\_header\_lines(.,values = c("Cuadro 2.1. Análisis exploratorio de variables cuantitativas"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020.") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(Media = function(x) format(round(x,2), big.mark = ""),  
 Desvío\_std= function(x) format(round(x, 2), big.mark = ""),   
 Mín= function(x) format(round(x,), big.mark = ""),   
 Q1= function(x) format(round(x,), big.mark = ""),   
 Mediana = function(x) format(round(x,), big.mark =""),  
 Q3 = function(x) format(round(x,), big.mark =""),  
 Máx = function(x) format(round(x,), big.mark =""))

| Cuadro 2.1. Análisis exploratorio de variables cuantitativas | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tipo | Variable | NA\_s | Media | Desvío\_std | Mín | Q1 | Mediana | Q3 | Máx | numeric.hist |
| numeric | Nro\_hogar | 0 | 1.06 | 1.22 | 1 | 1 | 1 | 1 | 61 | ▇▁▁▁▁ |
| numeric | Edad | 0 | 41.44 | 12.42 | 14 | 31 | 41 | 51 | 82 | ▃▇▇▃▁ |
| numeric | Total\_horas | 0 | 26.70 | 22.12 | 0 | 8 | 28 | 40 | 999 | ▇▁▁▁▁ |
| numeric | Ingreso | 0 | 23757.93 | 24437.94 | -9 | 6000 | 20000 | 35000 | 545000 | ▇▁▁▁▁ |
| Fuente: EPH 2º Trimestre 2020. | | | | | | | | | | |

A partir del Cuadro resúmen generado con la función *skim* del paquete *skirm* se observa que el dataset no cuenta con *missing values* (NA’s). Por otro lado, llama la atención el valor mínimo de la variable Ingreso (-9) y el máximo de la variable Total\_Horas (999). A partir del Diseño de Registro y Estructura de la base de la EPH se observa que el valor -9 corresponde a la no respuesta, mientras que el 999 indica la categoría de “No sabe/no responde”. Existen 1446 casos para el primero y sólo 1 para el segundo en el dataset. Por otro lado, 712 individuos reportaron un monto de ingreso de la ocupación principal de 0. Para el análisis no se incluirán los casos de no respuesta de ambas variables.

datos %>%   
 filter(Ingreso == -9) %>%   
 summarise(n = n())

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 1446

datos %>%   
 filter(Ingreso == 0) %>%   
 summarise(n = n())

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 712

datos %>%   
 filter(Total\_horas == 999) %>%   
 summarise(n = n())

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 1

## Características principales de las personas que no declararon ingresos

datos\_na <- datos %>%   
 filter(Ingreso == -9)

t\_rel <- datos\_na %>%   
 group\_by(Relacion) %>%   
 summarise(N = n())  
  
t\_rel <- t\_rel %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(N))  
  
t\_ed <- datos\_na %>%   
 group\_by(Nivel\_ed) %>%   
 summarise(N = n())  
  
t\_ed <- t\_ed %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(N))  
  
t\_cat <- datos\_na %>%   
 group\_by(Categoria\_ocup) %>%   
 summarise(N = n())  
  
t\_cat <- t\_cat %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(N))  
  
t\_inten <- datos\_na %>%   
 group\_by(Intensidad) %>%   
 summarise(N = n())  
  
t\_inten <- t\_inten %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(N))  
  
t\_estad <- datos\_na %>%   
 group\_by(Estado) %>%   
 summarise(N = n())  
  
t\_estad <- t\_estad %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(N))  
  
t\_sexo <- datos\_na %>%   
 group\_by(Sexo) %>%   
 summarise(N = n())  
  
t\_sexo <- t\_sexo %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(N))  
  
  
flextable((t\_rel),   
 col\_keys = names(t\_rel)) %>%   
 add\_header\_lines(.,values = c("Cuadro 2. Relación con el Jefe/a del hogar")) %>%  
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 2. Relación con el Jefe/a del hogar | | |
| --- | --- | --- |
| Relacion | N | Freq |
| Jefe/a | 664 | 45.92 % |
| Hijo/a Hijastro/a | 348 | 24.07 % |
| Conyuge / Pareja | 328 | 22.68 % |
| Yerno/Nuera | 35 | 2.42 % |
| Hermano/a | 22 | 1.52 % |
| Otros Familiares | 18 | 1.24 % |
| Nietro/a | 13 | 0.9 % |
| No familiares | 11 | 0.76 % |
| Madre/Padre | 7 | 0.48 % |
| Fuente: EPH 2º Trimestre 2020 | | |

flextable((t\_ed),   
 col\_keys = names(t\_ed)) %>%   
 add\_header\_lines(.,values = c("Cuadro 3. Nivel educativo") )%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 3. Nivel educativo | | |
| --- | --- | --- |
| Nivel\_ed | N | Freq |
| Superior universitaria completa | 448 | 30.98 % |
| Secundaria completa | 444 | 30.71 % |
| Superior universitaria incompleta | 203 | 14.04 % |
| Secundaria incompleta | 192 | 13.28 % |
| Primaria completa | 118 | 8.16 % |
| Primaria incompleta (incluye educacion especial) | 35 | 2.42 % |
| Sin instruccion | 6 | 0.41 % |
| Fuente: EPH 2º Trimestre 2020 | | |

flextable((t\_cat),   
 col\_keys = names(t\_cat)) %>%   
 add\_header\_lines(.,values = c("Cuadro 4. Categoría ocupacional")) %>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""))

| Cuadro 4. Categoría ocupacional | | |
| --- | --- | --- |
| Categoria\_ocup | N | Freq |
| Obrero o empleado | 1019 | 70.47 % |
| Cuenta propia | 362 | 25.03 % |
| Patron | 65 | 4.5 % |
| Fuente: EPH 2º Trimestre 2020 | | |

flextable((t\_inten),   
 col\_keys = names(t\_inten)) %>%   
 add\_header\_lines(.,values = c("Cuadro 5. Intensidad de la ocupación"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 5. Intensidad de la ocupación | | |
| --- | --- | --- |
| Intensidad | N | Freq |
| Ocupado pleno | 696 | 48.13 % |
| Sobreocupado | 312 | 21.58 % |
| Ocupado que no trabajo en la semana | 294 | 20.33 % |
| Subocupado por insuficiencia horaria | 144 | 9.96 % |
| Fuente: EPH 2º Trimestre 2020 | | |

flextable((t\_sexo),   
 col\_keys = names(t\_sexo)) %>%   
 add\_header\_lines(.,values = c("Cuadro 6. Sexo"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 6. Sexo | | |
| --- | --- | --- |
| Sexo | N | Freq |
| Varon | 854 | 59.06 % |
| Mujer | 592 | 40.94 % |
| Fuente: EPH 2º Trimestre 2020 | | |

A partir de los Cuadros (2, 3, 4, 5 y 6) es posible observar alguna de las características de las personas que no declararon ingresos de su ocupación principal. En primer lugar, aproximadamente el 46% de las personas se reconocen como jefe/a del hogar, en segundo se percibe que aquellas personas con mayor nivel educativo tienden, en promedio, a no declarar su ingreso (casi el 62% de las personas cuenta con secundaria o universitario completo). En tercero, el 70,47% resultan empleados o obreros mientras que aproximadamente la mitad de ellos (el 48,13%) son ocupados plenos. Por último, se observa una tendencia a que los varones sean más reticentes a declarar su ingreso en relación a la mujeres.

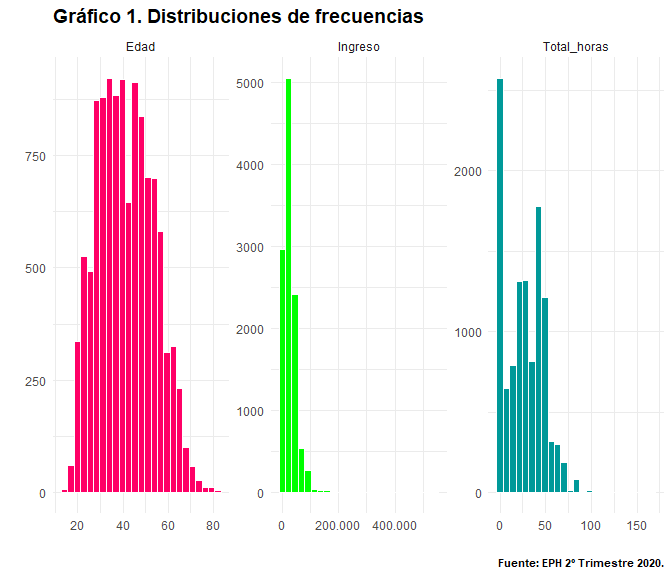
## Resúmen numérico para las variables cuantitativas

datos\_conti\_f <- datos %>%   
 filter(Ingreso!=-9, Total\_horas!=999) %>%   
 dplyr::select(Total\_horas, Ingreso, Edad)  
   
datos\_conti\_r <- skim(datos\_conti\_f) %>%   
 rename(Variable = skim\_variable) %>%   
 dplyr::select(-complete\_rate, -n\_missing, -numeric.hist, -skim\_type) %>%   
 mutate(across(where(is.numeric),round, 2)) %>%   
 rename(Media = numeric.mean,   
 Desvío\_std=numeric.sd,   
 Mín=numeric.p0,  
 Q1=numeric.p25,  
 Mediana=numeric.p50,  
 Q3=numeric.p75,  
 Máx=numeric.p100)   
  
  
datos\_conti\_b <- datos\_conti\_f %>%   
 pivot\_longer(cols = c(names(datos\_conti\_f))) %>%   
 group\_by(name) %>%   
 summarise(media = mean(value),  
 desvio = sd(value),  
 mediana = median(value),   
 CV = round(desvio/media\*100,2),  
 Asimetría = round(3\*(media-mediana)/desvio,2)) %>%   
 dplyr::select(-media, -desvio, -mediana)  
   
  
datos\_conti\_r <- left\_join(datos\_conti\_r, datos\_conti\_b, by = c("Variable" = "name"))   
  
flextable((datos\_conti\_r),   
 col\_keys = c("Variable", "Media", "Desvío\_std", "Mín", "Q1", "Mediana", "Q3", "Máx", "CV", "Asimetría")) %>%   
 add\_header\_lines(.,values = c("Cuadro 2. Medidas descriptivas: tendencia central, dispersión y asimetría")) %>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(Media = function(x) format(round(x, 3), big.mark = ""),  
 Desvío\_std= function(x) format(round(x, 3), big.mark = ""),   
 Q1= function(x) format(round(x, 3), big.mark = ""),   
 Mediana = function(x) format(round(x, 3), big.mark =""),  
 Q3=function(x) format(round(x,3), big.mark=""),  
 Máx=function(x) format(round(x,3), big.mark=""))

| Cuadro 2. Medidas descriptivas: tendencia central, dispersión y asimetría | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Media | Desvío\_std | Mín | Q1 | Mediana | Q3 | Máx | CV | Asimetría |
| Total\_horas | 26.48 | 20.37 | 0 | 7 | 27 | 40 | 168 | 76.91 | -0.08 |
| Ingreso | 26786.35 | 24334.39 | 0 | 10000 | 23000 | 36000 | 545000 | 90.85 | 0.47 |
| Edad | 41.32 | 12.36 | 14 | 31 | 41 | 50 | 82 | 29.91 | 0.08 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | | | |

## Distribución de frecuencias

options(scipen = 99)  
datos\_conti\_f %>%   
 pivot\_longer(cols = c(names(datos\_conti\_f))) %>%   
   
ggplot()+   
 geom\_histogram(aes(x = value, fill = name), bins=25, color = "white", show.legend = F)+  
 facet\_wrap(~name, scales='free', ncol =3) +   
 scale\_x\_continuous(labels = scales::number\_format(big.mark = ".",decimal.mark = ","))+  
 scale\_fill\_manual(values = c("#FF0066", "#00FF00", "#009999")) +  
 labs(x = " ",  
 y = " ",  
 title = "Gráfico 1. Distribuciones de frecuencias ",  
 caption = "Fuente: EPH 2º Trimestre 2020.") + theme\_minimal() +  
 theme(plot.title = element\_text(size = 14, face = "bold"),  
 plot.caption = element\_text(size=8, face = "bold"))



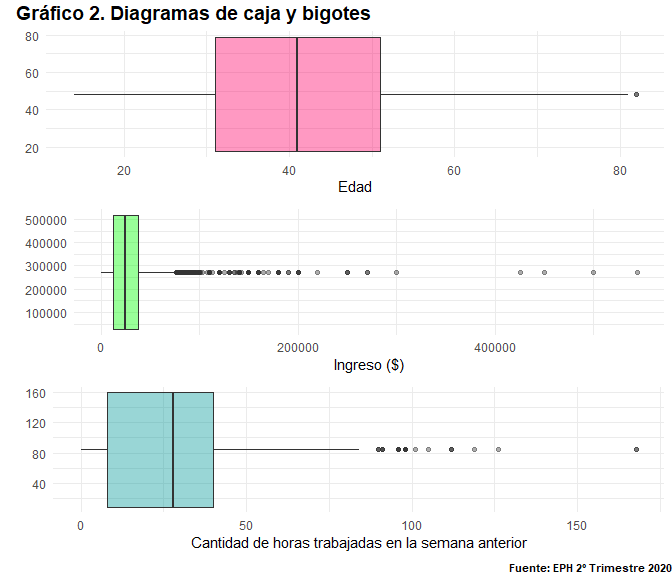
ggsave("graficos/histogramas.png")

## Saving 7 x 6 in image

## Boxplots

a <- ggplot(data = datos, aes(x=Edad, y=Edad)) +   
 geom\_boxplot(fill = "#FF0066", alpha =0.4) +   
 labs (x = " ",   
 y = "Edad") + coord\_flip() + theme\_minimal()  
  
b <- ggplot(data = filter(datos, Ingreso>0), aes(Ingreso, y = Ingreso)) +   
 geom\_boxplot(fill = "#00FF00", alpha=0.4) +   
 labs (x = " ",  
 y = "Ingreso ($)")+ coord\_flip() + theme\_minimal()  
  
c <- ggplot(data = filter(datos, Total\_horas != 999), aes(x=Total\_horas, y=Total\_horas)) +   
 geom\_boxplot(fill = "#009999", alpha = 0.4) +   
 labs (x = " ",   
 y = "Cantidad de horas trabajadas en la semana anterior") + coord\_flip() +   
 theme\_minimal()  
  
boxplots <- grid.arrange(a,b,c, ncol=1, nrow=3, widths = c(8), heights = c(5, 5,5),  
 top = textGrob("Gráfico 2. Diagramas de caja y bigotes",  
 gp=gpar(fontsize=14, fontface = "bold"),  
 hjust=0.9),  
 bottom = textGrob(  
 "Fuente: EPH 2º Trimestre 2020",  
 gp = gpar(fontface = "bold", fontsize = 8),  
 hjust = 1,  
 x = 1))

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?  
  
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?  
  
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



ggsave("graficos/boxplots.jpg", boxplots)

## Saving 7 x 6 in image

A partir del Cuadro 1 y de los Gráficos 1 y 2, es posible observar cómo se distribuyen las variables continuas del dataset. En promedio, las personas relevadas en la encuesta (ocupadas) poseen un ingreso de su ocupación principal de 26786.35 pesos, una edad media de 41 años y han trabajado en la última semana 27 horas aproximadamente. Con respecto a la dispersión se aprecia, a partir de la estimación del Coeficiente de Variación (CV) y de los gráficos (histogramas y diagramas de caja y bigote), que la variable Edad posee una menor variabilidad en torno a su media en comparación al resto de las variables (presenta el menor CV). Por otro lado, dicha variable resulta ser la más simétrica de las 3, aspecto que se observa, de forma analítica, al comparar el Coeficiente de Asimetría de Pearson estimado (el mismo es bastante cercano a 0) y, a su vez, de forma visual, a través de los histogramas (la distribución de dicha variable tiende a asimilarse a una campana) y en el diagrama de caja y bigote se observa que el largo del bigote inferior y superior es bastante similar.

La distribución de la variable Ingreso resulta ser la más asimétrica (registra una marcada asimetría positiva), en tanto posee un Coeficiente de Asimetría de Pearson mayor a 0 (0.47) y que resulta superior al resto de las distribuciones, lo que indica que la gran mayoría de las personas relevadas declararon ingresos menores a la media (subpromediales). En efecto, el 75% de las personas registraron ingresos de la ocupación principal de hasta 36000 pesos (Cuadro 1). Este aspecto también se corrobora al observar tanto el histograma como el diagrama de caja y bigote. La variable Total\_horas presenta una leve asimetría negativa, es decir que, en principio, la mayoría de las personas trabajaron durante la última semana más que la media.

Un aspecto que se destaca del análisis exploratorio resulta la existencia de datos *atípicos* en las variables Ingreso y Total\_horas. Tal como se destacó con anterioridad, el 75% de las personas registraron ingresos de hasta 36000 pesos, mientras que el restante 25% posee ingresos que oscilan entre dicho valor y 545000 pesos, lo que denota la gran variabilidad que existe en este último segmento de la muestra. Para los análisis y visualizaciones posteriores se aplicará el logoritmo en base 10 a dicha variable. Dicha transformación puede utilizarse cuando existen grandes disparidades en una variable, porque permite que los valores pequeños tengan espacio para diferenciarse, y a la vez que los muy grandes no aparezcan tan alejados, logrando, por lo general, visualizaciones más compactas y legibles.

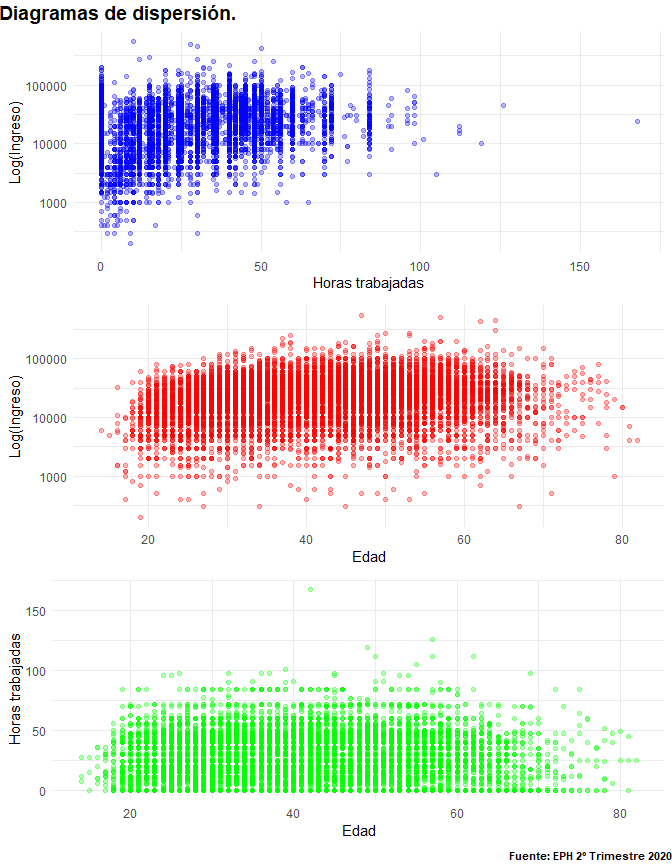
Con respecto a la variable Total\_horas, el 75% de las personas relevadas trabajó en el última semana hasta 40 horas, el segmento restante fluctuó entre 40 y 168 horas. Este último valor representaría a una persona que trabajó en promedio 24 horas diarias durante 7 días. Dado que este dato puede resultar anómalo, es posible observar el caracter de la ocupación principal de las personas basado en el Clasificador Nacional de Ocupaciones (CNO), tal como se aprecia en el Cuadro 3. Existen 2 personas que declararon haber trabajado 168 horas en la última semana, las mismas se desempeñan en servicios policiales y en cuidado y atención de personas, actividades en las cuáles pueden darse situaciones en donde se trabajé en formato de guardia.

CNO <- datos %>%   
 filter(Total\_horas != 999) %>%   
 arrange(desc(Total\_horas)) %>%   
 dplyr::select(Categoria, Total\_horas) %>%   
 head(n=10)  
  
flextable(CNO,   
 cwidth=2, cheight = 1) %>%   
 add\_header\_lines(.,values = c("Cuadro 3. Categoría ocupacional de las 10 personas que más trabajaron en la semana anterior")) %>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 3. Categoría ocupacional de las 10 personas que más trabajaron en la semana anterior | |
| --- | --- |
| Categoria | Total\_horas |
| Ocupaciones de servicios policiales | 168 |
| Ocupaciones del cuidado y la atención de personas | 168 |
| Ocupaciones del transporte | 126 |
| Ocupaciones de la comercialización directa | 119 |
| Ocupaciones del cuidado y la atención de personas | 112 |
| Ocupaciones de la producción ganadera | 112 |
| Ocupaciones de la comercialización directa | 112 |
| Ocupaciones de la comercialización directa | 105 |
| Directivos de pequeñas y microempresas | 101 |
| Directivos de medianas empresas privadas productoras de bienes y servicios | 98 |
| Fuente: EPH 2º Trimestre 2020 | |

## Diagramas de dispersión

g1 <- ggplot(data = filter(datos\_conti\_f, Ingreso !=0), aes(x= Total\_horas, y=Ingreso)) +  
 geom\_point(alpha = 0.3, colour = "blue", show.legend = F) +   
 labs(x = "Horas trabajadas",  
 y = "Log(Ingreso)")+ scale\_y\_log10() + theme\_minimal()   
  
g2 <- ggplot(data = filter(datos\_conti\_f, Ingreso !=0), aes(x= Edad, y=Ingreso)) +  
 geom\_point(alpha = 0.3, colour = "red", show.legend = F) +   
 labs(x = "Edad",  
 y = "Log(Ingreso)")+ scale\_y\_log10() + theme\_minimal()   
  
g3 <- ggplot(datos\_conti\_f, aes(x= Edad, y=Total\_horas)) +  
 geom\_point(alpha = 0.3, colour = "green", show.legend = F) +   
 labs(x = "Edad",  
 y = "Horas trabajadas")+theme\_minimal()  
  
g <- grid.arrange(g1,g2,g3, ncol=1, nrow=3,  
 top = textGrob("Gráfico 3. Diagramas de dispersión.",  
 gp=gpar(fontsize=14, fontface = "bold"),  
 hjust = 1.3),  
 bottom = textGrob(  
 "Fuente: EPH 2º Trimestre 2020",  
 gp = gpar(fontface="bold", fontsize = 8),  
 hjust = 1,  
 x = 1))

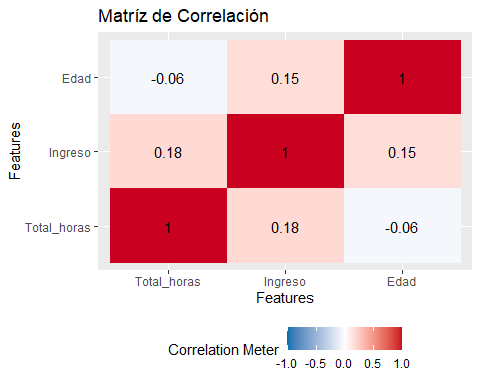


ggsave("graficos/dispersion.jpg", g)

## Saving 7 x 9 in image

## Matríz de Correlación

# datos\_conti\_f <- datos\_conti\_f %>%   
# mutate(log\_ingreso = log(Ingreso))  
  
m <- plot\_correlation(datos\_conti\_f, title = "Matríz de Correlación",  
 theme\_config = list(legend.position = "bottom", axis.text.x = NULL))



ggsave("graficos/matriz.jpg", m)

## Saving 5 x 4 in image

## Test de Correlación

cor\_1 <- cor.test(datos\_conti\_f$Ingreso, datos\_conti\_f$Total\_horas)  
estimate\_1 <- round(cor\_1$estimate,2)  
p\_value\_1 <- cor\_1$p.value  
int\_1 <- round(cor\_1$conf.int,2)  
  
cor\_2 <- cor.test(datos\_conti\_f$Ingreso, datos\_conti\_f$Edad)  
estimate\_2 <- round(cor\_2$estimate,2)  
p\_value\_2 <- cor\_2$p.value  
int\_2 <- round(cor\_2$conf.int,2)  
  
cor\_3 <- cor.test(datos\_conti\_f$Total\_horas, datos\_conti\_f$Edad)  
estimate\_3 <- round(cor\_3$estimate,2)  
p\_value\_3 <- cor\_3$p.value  
int\_3 <- round(cor\_3$conf.int,2)  
  
  
tabla\_corr <- data.frame(matrix(ncol = 4, nrow = 3))  
rownames(tabla\_corr)=c("Ingreso - Total\_horas","Ingreso - Edad","Edad - Total\_horas")  
colnames(tabla\_corr)=c("Estimado", "Valor\_p", "Lim. inferior", "Lim.superior")  
  
tabla\_corr <- cbind(Variables = rownames(tabla\_corr), tabla\_corr)  
rownames(tabla\_corr) <- NULL  
  
  
tabla\_corr[1,2]=estimate\_1  
tabla\_corr[2,2]=estimate\_2  
tabla\_corr[3,2]=estimate\_3  
tabla\_corr[1,3]=p\_value\_1  
tabla\_corr[2,3]=p\_value\_2  
tabla\_corr[3,3]=p\_value\_3  
tabla\_corr[1,4]=int\_1[1]  
tabla\_corr[2,4]=int\_2[1]  
tabla\_corr[3,4]=int\_3[1]  
tabla\_corr[1,5]=int\_1[2]  
tabla\_corr[2,5]=int\_2[2]  
tabla\_corr[3,5]=int\_3[2]  
  
  
flextable((tabla\_corr),  
 col\_keys = names(tabla\_corr)) %>%   
 add\_header\_lines(.,values = c("Cuadro 4. Test de correlaciones")) %>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(Valor\_p = function(x) format(round(x, 9), big.mark = ""))

| Cuadro 4. Test de correlaciones | | | | |
| --- | --- | --- | --- | --- |
| Variables | Estimado | Valor\_p | Lim. inferior | Lim.superior |
| Ingreso - Total\_horas | 0.18 | 0.000000000 | 0.16 | 0.20 |
| Ingreso - Edad | 0.15 | 0.000000000 | 0.14 | 0.17 |
| Edad - Total\_horas | -0.06 | 0.000000004 | -0.07 | -0.04 |
| Fuente: EPH 2º Trimestre 2020 | | | | |

Los análisis de correlación indican que las 3 relaciones entre pares de las variables continuas del dataset son significativas, los valores p asociados a cada relación son menores a un nivel de significancia del 5% (Cuadro 4). De dichas relaciones la correspondiente a Ingreso y Total\_horas presenta mayor intensidad (lineal positiva), aunque cabe destacar que la misma es bastante leve (0.18 es el Coeficiente de Correlacion estimado). A partir Gráfico 3 se observa que la gran mayoría de las personas trabajaron hasta 50 horas semanales, pero se percibe una gran variabilidad en torno al logaritmo del ingreso percibido en ese segmento. A su vez, es posible registrar los datos extremos de 168 horas trabajadas identificadas en el Cuadro 3.

Con respecto a las otras correlaciones, se observa que la Edad presenta una menor intensidad en su relación con el ingreso, y que la misma con las horas trabajadas es negativa, lo que indicaría que a medida que las personas crecen tienden a trabajar menos horas.

Para complementar el análisis de correlación se estimará la *Variance Inflation Factors* (VIF) para corroborar la posible existencia o no de multicolinealidad, según la literatura un VIF mayor a 5 o a 10 estaría indicando la presencia de multicolinealidad. La VIF para el predictor *j* se estima a partir de la siguiente fórmula:

=

Siendo: el Coeficiente de Determinación obtenido de la regresión del predictor *j* con el resto de los predictores. En este caso, al tener sólo 2 predictores cuantitativos continuos se obtendrá un único VIF.

## Variance Inflation Factors

#Regresión entre Total\_horas ~ Edad  
r\_cuadrado <- round(summary(lm(Total\_horas ~ Edad, data = datos\_conti\_f))$r.squared,6)  
  
VIF <- 1/(1-r\_cuadrado)  
  
tabla\_vif <- data.frame(matrix(ncol = 1, nrow = 2))  
rownames(tabla\_vif)=c("Total\_horas","Edad")  
colnames(tabla\_vif)=c("VIF")  
  
tabla\_vif <- cbind(Variables = rownames(tabla\_vif), tabla\_vif)  
rownames(tabla\_vif) <- NULL  
  
  
tabla\_vif[1,2]=round(VIF,4)  
tabla\_vif[2,2]=round(VIF,4)  
  
  
flextable((tabla\_vif),  
 col\_keys = names(tabla\_vif)) %>%   
 add\_header\_lines(.,values = c("Cuadro 5. Variance Inflation Factors")) %>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit")

| Cuadro 5. Variance Inflation Factors | |
| --- | --- |
| Variables | VIF |
| Total\_horas | 1.0031 |
| Edad | 1.0031 |
| Fuente: EPH 2º Trimestre 2020 | |

Los valores de VIF obtenidos en el Cuadro 5, sumado a la Matríz y Test de Correlación estimados, indicarían que no habría problemas de colinealidad y multicolinealidad en el dataset.

## Análisis de las variables categóricas

## Recodificación

Se decidió recodificar ciertas variables del dataset para una mejor visualización y modelización de los datos

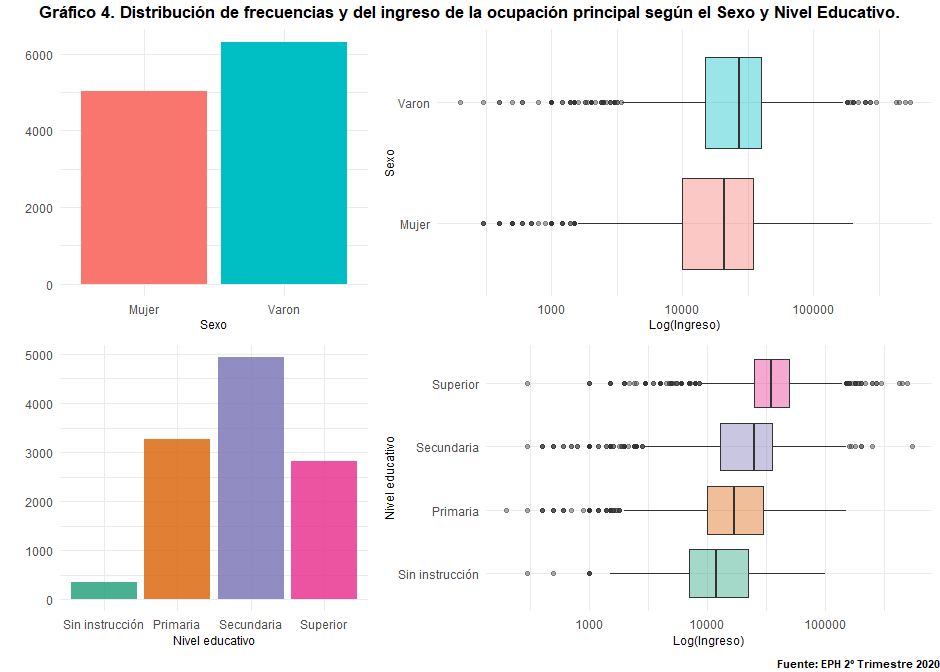
#Niveles de la variable nivel educativo  
table(datos$Nivel\_ed)

##   
## Primaria completa   
## 1540   
## Primaria incompleta (incluye educacion especial)   
## 357   
## Secundaria completa   
## 3804   
## Secundaria incompleta   
## 2033   
## Sin instruccion   
## 29   
## Superior universitaria completa   
## 3258   
## Superior universitaria incompleta   
## 1782

#Recodificamos la variable de nivel educativo  
datos<- datos %>%   
 mutate(Nivel\_ed = case\_when(Nivel\_ed %in% c("Primaria incompleta (incluye educacion especial)", "Sin instruccion") ~ "Sin instrucción",  
 Nivel\_ed %in% c("Primaria completa", "Secundaria incompleta") ~ "Primaria",  
 Nivel\_ed %in% c("Secundaria completa","Superior universitaria incompleta") ~ "Secundaria",  
 Nivel\_ed %in% c("Superior universitaria completa") ~ "Superior"))  
  
datos$Nivel\_ed <- as.factor(datos$Nivel\_ed)  
datos$Nivel\_ed <- factor(datos$Nivel\_ed, levels = c("Sin instrucción","Primaria","Secundaria","Superior"))  
  
#Recodificamos la variable Estado, Intensidad, Relación y Tipo  
datos<- datos %>%   
 mutate(Estado = case\_when(Estado == "casado?" ~ "Casado",  
 Estado == "soltero/a?" ~ "soltero/a",  
 Estado == "unido?" ~ "unido",  
 Estado == "viudo/a?" ~ "viudo/a",  
 Estado == "separado/a o divorsiado/a?" ~ "separado/a"),  
 Intensidad = case\_when(Intensidad == "Ocupado que no trabajo en la semana" ~ "No\_trabajo",  
 Intensidad == "Subocupado por insuficiencia horaria" ~ "Subocupado",  
 TRUE ~ Intensidad),  
  
 Relacion = case\_when(Relacion == "Conyuge / Pareja" ~ "Conyuge",  
 Relacion == "Hijo/a Hijastro/a" ~ "Hijo/a",  
 Relacion == "Nietro/a" ~ "Nieto/a",  
 Relacion %in% c("Otros Familiares", "Suegro/a", "Yerno/Nuera") ~ "Otros",  
 Relacion == "No familiares" ~ "No\_fliar",  
 TRUE ~ Relacion),  
 Tipo = case\_when(Tipo == "...privada" ~ "Privado",  
 Tipo == "...estatal" ~ "Estatal",  
 Tipo == "...de otro tipo" ~ "Otro",  
 TRUE ~ Tipo))

## Gráfico múltiple de distribución del ingreso y frecuencias: Sexo y Nivel educativo

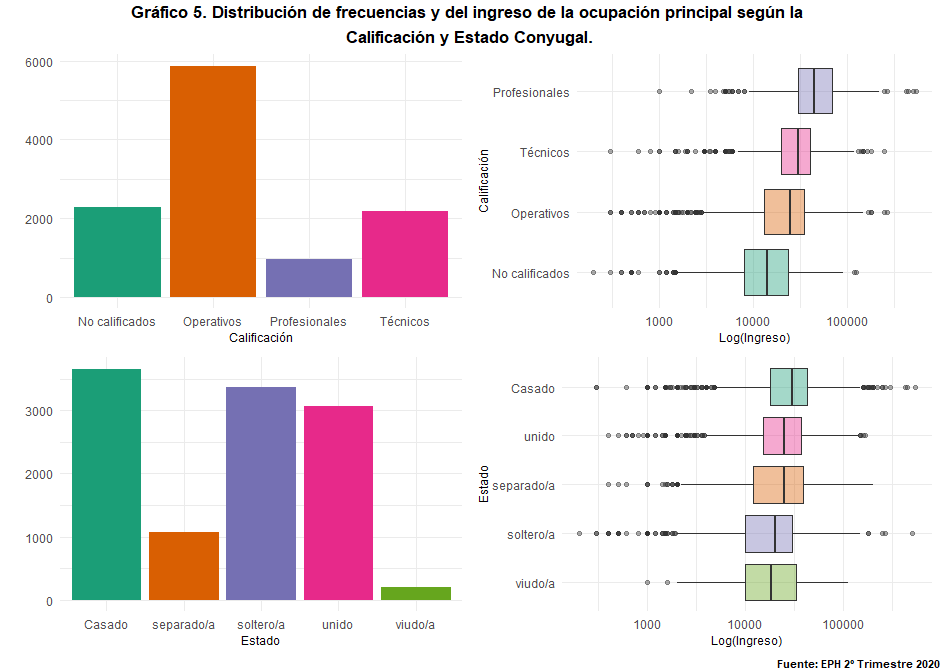
datos\_f <- datos %>%   
 filter(Ingreso >=0, Total\_horas!=999)  
  
g4\_1 <- ggplot(data = datos\_f) + geom\_bar(aes(x= Sexo, fill = Sexo), show.legend = F) +   
 labs (x = "Sexo",  
 y = " ") + theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
g4 <- ggplot(data = datos\_f,aes(x=reorder(Sexo,Ingreso, FUN = median), y=Ingreso, fill = Sexo)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +   
 labs (x = "Sexo",  
 y = "Log(Ingreso)") + theme\_minimal() + theme(axis.title.x = element\_text(size = 9),  
 axis.title.y = element\_text(size = 9))  
  
g5\_1 <- ggplot(data = datos\_f) + geom\_bar(aes(x= Nivel\_ed, fill = Nivel\_ed), show.legend = F, alpha = 0.8) +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Nivel educativo",  
 y = " ") + theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
  
g5 <- ggplot(data = datos\_f, aes(x=reorder(Nivel\_ed,Ingreso, FUN = median), y=Ingreso, fill = Nivel\_ed)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Nivel educativo",  
 y = "Log(Ingreso)")+  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9),  
 axis.title.y = element\_text(size = 9))  
  
gm <- grid.arrange(g4\_1,g4,g5\_1,g5, ncol=2, nrow=2, widths = c(6, 9), heights = c(7, 7),  
 top = textGrob("Gráfico 4. Distribución de frecuencias y del ingreso de la ocupación principal según el Sexo y Nivel Educativo.",  
 gp=gpar(fontsize=13, fontface = "bold"),  
 hjust = 0.5),  
 bottom = textGrob(  
 "Fuente: EPH 2º Trimestre 2020",  
 gp = gpar(fontface = "bold", fontsize = 8),  
 hjust = 1,  
 x = 1))



ggsave("graficos/g\_multiple1.jpg", gm)

## Gráfico múltiple de distribución del ingreso y frecuencias: Calificación y Estado

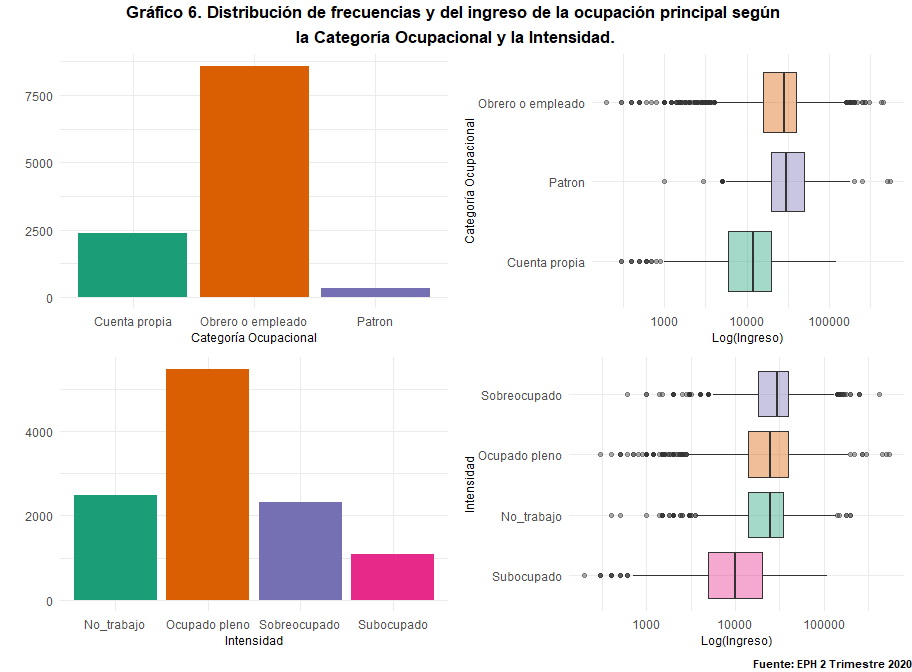
g6\_1 <- ggplot(data = filter(datos\_f, Calificacion %in% c("No calificados", "Operativos", "Profesionales", "Técnicos"))) + geom\_bar(aes(x= Calificacion, fill = Calificacion), show.legend = F) + scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Calificación",  
 y = " ") +  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
  
g6 <- ggplot(data = filter(datos\_f, Calificacion %in% c("No calificados", "Operativos", "Profesionales", "Técnicos"), ), aes(x=reorder(Calificacion,Ingreso, FUN = median), y=Ingreso, fill = Calificacion)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Calificación",  
 y = "Log(Ingreso)")+  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9),  
 axis.title.y = element\_text(size = 9))  
  
  
g7\_1 <- ggplot(data = datos\_f) + geom\_bar(aes(x= Estado, fill = Estado), show.legend = F) +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Estado",  
 y = " ") +  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
  
  
g7 <- ggplot(data = datos\_f, aes(x=reorder(Estado,Ingreso, FUN = median), y=Ingreso, fill = Estado)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Estado",  
 y = "Log(Ingreso)") +  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9),  
 axis.title.y = element\_text(size = 9))  
  
  
  
gm\_2 <- grid.arrange(g6\_1,g6,g7\_1,g7, ncol=2, nrow=2, widths = c(8, 8), heights = c(6, 6),  
 top = textGrob("Gráfico 5. Distribución de frecuencias y del ingreso de la ocupación principal según la   
Calificación y Estado Conyugal.",  
 gp=gpar(fontsize=13, fontface = "bold"),  
 hjust = 0.5),  
 bottom = textGrob(  
 "Fuente: EPH 2º Trimestre 2020",  
 gp = gpar(fontface = "bold", fontsize = 8),  
 hjust = 1,  
 x = 1))



ggsave("graficos/g\_multiple2.jpg", gm\_2)

## Gráfico múltiple de distribución del ingreso y frecuencias: Categoría ocupacional e Intensidad

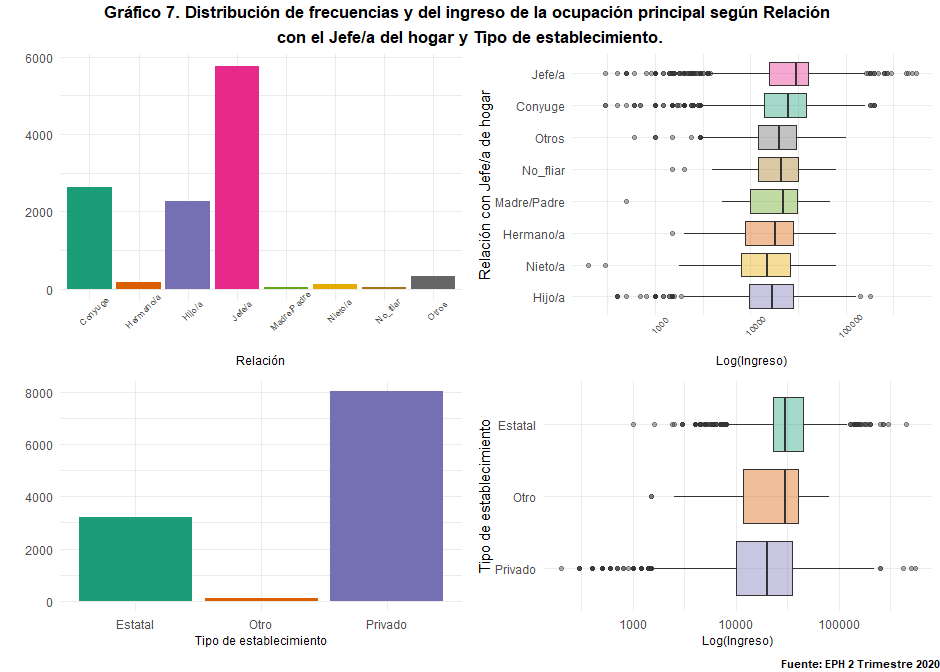
g8\_1 <- ggplot(data = filter(datos\_f, Categoria\_ocup %in% c("Cuenta propia", "Obrero o empleado", "Patron"))) +   
geom\_bar(aes(x= Categoria\_ocup, fill = Categoria\_ocup), show.legend = F) + scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Categoría Ocupacional",  
 y = " ") +  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
  
g8 <- ggplot(data = filter(datos, Categoria\_ocup %in% c("Cuenta propia", "Obrero o empleado", "Patron")), aes(x=reorder(Categoria\_ocup,Ingreso, FUN = median), y=Ingreso, fill = Categoria\_ocup)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +  
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Categoría Ocupacional",  
 y = "Log(Ingreso)")+  
 theme\_minimal() + theme(axis.title.y = element\_text(size = 9),  
 axis.title.x = element\_text(size = 9))  
  
g9\_1 <- ggplot(data = datos\_f) +   
geom\_bar(aes(x= Intensidad, fill = Intensidad), show.legend = F) + scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Intensidad",  
 y = " ")+  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
g9 <- ggplot(data = datos, aes(x=reorder(Intensidad,Ingreso, FUN=median), y=Ingreso, fill = Intensidad)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Intensidad",  
 y = "Log(Ingreso)")+  
 theme\_minimal() + theme(axis.title.y = element\_text(size = 9),  
 axis.title.x = element\_text(size = 9))  
  
gm\_3 <- grid.arrange(g8\_1,g8,g9\_1,g9, ncol=2, nrow=2, widths = c(8, 8), heights = c(6, 6),  
 top = textGrob("Gráfico 6. Distribución de frecuencias y del ingreso de la ocupación principal según   
la Categoría Ocupacional y la Intensidad.",  
 gp=gpar(fontsize=13, fontface = "bold"),  
 hjust = 0.5),  
 bottom = textGrob(  
 "Fuente: EPH 2 Trimestre 2020",  
 gp = gpar(fontface = "bold", fontsize = 8),  
 hjust = 1,  
 x = 1))



ggsave("graficos/g\_multiple3.jpg", gm\_3)

## Gráfico múltiple de distribución del ingreso y frecuencias: Relación con el Jefe/a del hogar y Tipo de establecimiento

g10\_1 <- ggplot(data = datos\_f) +   
geom\_bar(aes(x= Relacion, fill = Relacion), show.legend = F) + scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Relación",  
 y = " ") +  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9),  
 axis.text.x = element\_text(size =7,angle = 45))  
  
  
g10 <- ggplot(data = datos, aes(x=reorder(Relacion, Ingreso, FUN = median), y=Ingreso, fill = Relacion)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +  
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Relación con Jefe/a de hogar",  
 y = "Log(Ingreso)") +   
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9),  
 axis.text.x = element\_text(size =7,angle = 45))  
  
  
g11\_1 <- ggplot(data = datos\_f) +   
geom\_bar(aes(x= Tipo, fill = Tipo), show.legend = F) + scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Tipo de establecimiento",  
 y = " ") +  
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
  
  
g11 <- ggplot(data = filter(datos, Tipo %in% c("Privado", "Estatal", "Otro")), aes(x=reorder(Tipo,Ingreso, FUN =median), y=Ingreso, fill = Tipo)) +   
 geom\_boxplot(show.legend = F, alpha = 0.4) + scale\_y\_log10() + coord\_flip() +   
 scale\_fill\_brewer(type = "qual", palette = "Dark2") +  
 labs (x = "Tipo de establecimiento",  
 y = "Log(Ingreso)") +   
 theme\_minimal() + theme(axis.title.x = element\_text(size = 9))  
  
  
gm\_4 <- grid.arrange(g10\_1,g10,g11\_1,g11, ncol=2, nrow=2, widths = c(8, 8), heights = c(7, 6),  
 top = textGrob("Gráfico 7. Distribución de frecuencias y del ingreso de la ocupación principal según Relación   
con el Jefe/a del hogar y Tipo de establecimiento.",  
 gp=gpar(fontsize=13, fontface ="bold"),  
 hjust = 0.5),  
 bottom = textGrob(  
 "Fuente: EPH 2 Trimestre 2020",  
 gp = gpar(fontface = "bold", fontsize = 8),  
 hjust = 1,  
 x = 1))



ggsave("graficos/g\_multiple4.jpg", gm\_4)

## Análisis de la tendencia central, dispersión y asimetría del ingreso de la ocupación principal en función de las variables categóricas

tabla\_sexo <- datos\_f %>%   
 group\_by(Sexo) %>%   
 summarise(N = n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_sexo <- tabla\_sexo %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Sexo, N, Freq, everything())  
  
  
flextable((tabla\_sexo),   
 col\_keys = names(tabla\_sexo)) %>%   
 add\_header\_lines(.,values = c("Cuadro 6. Distribución del ingreso de la ocupación principal según Sexo"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x, 2), big.mark = ""),   
 Mediana= function(x) format(round(x, 2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x, 2), big.mark =""))

| Cuadro 6. Distribución del ingreso de la ocupación principal según Sexo | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sexo | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Varon | 6318 | 55.64 % | 29372.35 | 25000 | 26998.10 | 91.92 | 0.49 |
| Mujer | 5038 | 44.36 % | 23543.33 | 20000 | 20047.23 | 85.15 | 0.53 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

tabla\_nivel\_Ed <- datos\_f %>%   
 group\_by(Nivel\_ed) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_nivel\_Ed <- tabla\_nivel\_Ed %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Nivel\_ed, N, Freq, everything())  
  
  
flextable((tabla\_nivel\_Ed),   
 col\_keys = names(tabla\_nivel\_Ed)) %>%   
 add\_header\_lines(.,values = c("Cuadro 7. Distribución del ingreso de la ocupación principal según Nivel Educativo"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x, 2), big.mark = ""),   
 Mediana= function(x) format(round(x, 2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x, 2), big.mark =""))

| Cuadro 7. Distribución del ingreso de la ocupación principal según Nivel Educativo | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Nivel\_ed | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Superior | 2809 | 24.74 % | 39442.84 | 35000 | 31648.63 | 80.24 | 0.42 |
| Secundaria | 4939 | 43.49 % | 25586.95 | 22800 | 21369.89 | 83.52 | 0.39 |
| Primaria | 3263 | 28.73 % | 19009.37 | 15000 | 16441.15 | 86.49 | 0.73 |
| Sin instrucción | 345 | 3.04 % | 14461.74 | 10000 | 14037.70 | 97.07 | 0.95 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

tabla\_calif <- datos\_f %>%   
 filter(Calificacion %in% c("No calificados", "Operativos", "Profesionales", "Técnicos")) %>%   
 group\_by(Calificacion) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_calif <- tabla\_calif %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Calificacion, N, Freq, everything())  
  
  
flextable((tabla\_calif),   
 col\_keys = names(tabla\_calif)) %>%   
 add\_header\_lines(.,values = c("Cuadro 8. Distribución del ingreso de la ocupación principal según la calificación del puesto de trabajo"))%>%  
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x, 2), big.mark = ""),   
 Mediana= function(x) format(round(x, 2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x, 2), big.mark =""))

| Cuadro 8. Distribución del ingreso de la ocupación principal según la calificación del puesto de trabajo | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Calificacion | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Profesionales | 963 | 8.54 % | 52438.71 | 45000 | 45976.63 | 87.68 | 0.49 |
| Técnicos | 2169 | 19.24 % | 32334.30 | 30000 | 20640.23 | 63.83 | 0.34 |
| Operativos | 5865 | 52.04 % | 24792.50 | 22000 | 19535.31 | 78.80 | 0.43 |
| No calificados | 2274 | 20.18 % | 15347.40 | 12000 | 12656.16 | 82.46 | 0.79 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

#la suma de Freq no da exactamente 1 por cuestión de redondeo

tabla\_estado <- datos\_f %>%   
 group\_by(Estado) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_estado <- tabla\_estado %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Estado, N, Freq, everything())  
  
  
flextable((tabla\_estado),   
 col\_keys = names(tabla\_estado)) %>%   
 add\_header\_lines(.,values = c("Cuadro 9. Distribución del ingreso de la ocupación principal según el Estado Conyugal"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x, 2), big.mark = ""),   
 Mediana= function(x) format(round(x, 2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x, 2), big.mark =""))

| Cuadro 9. Distribución del ingreso de la ocupación principal según el Estado Conyugal | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Estado | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Casado | 3656 | 32.19 % | 32470.92 | 30000 | 29155.42 | 89.79 | 0.25 |
| unido | 3068 | 27.02 % | 26340.15 | 23000 | 20629.93 | 78.32 | 0.49 |
| separado/a | 1072 | 9.44 % | 26295.58 | 22000 | 22663.27 | 86.19 | 0.57 |
| viudo/a | 196 | 1.73 % | 22757.91 | 17000 | 19958.36 | 87.70 | 0.87 |
| soltero/a | 3364 | 29.62 % | 21406.38 | 18000 | 20867.97 | 97.48 | 0.49 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

tabla\_cat\_ocup <- datos\_f %>%   
 filter(Categoria\_ocup %in% c("Cuenta propia", "Obrero o empleado", "Patron")) %>%   
 group\_by(Categoria\_ocup) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_cat\_ocup <- tabla\_cat\_ocup %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 dplyr::select(Categoria\_ocup, N, Freq, everything())  
  
  
flextable((tabla\_cat\_ocup),   
 col\_keys = names(tabla\_cat\_ocup)) %>%   
 add\_header\_lines(.,values = c("Cuadro 10. Distribución del ingreso de la ocupación principal según la Categoría ocupacional"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x,2), big.mark = ""),   
 Mediana= function(x) format(round(x,2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x,2), big.mark =""))

| Cuadro 10. Distribución del ingreso de la ocupación principal según la Categoría ocupacional | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Categoria\_ocup | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Cuenta propia | 2364 | 20.95 % | 14044.99 | 10000 | 15086.03 | 107.41 | 0.80 |
| Obrero o empleado | 8583 | 76.07 % | 30152.80 | 27000 | 23523.02 | 78.01 | 0.40 |
| Patron | 336 | 2.98 % | 36255.65 | 25000 | 51076.86 | 140.88 | 0.66 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

tabla\_intensidad <- datos\_f %>%   
 group\_by(Intensidad) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_intensidad <- tabla\_intensidad %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Intensidad, N, Freq, everything())  
  
  
flextable((tabla\_intensidad),   
 col\_keys = names(tabla\_intensidad)) %>%   
 add\_header\_lines(.,values = c("Cuadro 11. Distribución del ingreso de la ocupación principal según la Intensidad de la ocupación"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x,2), big.mark = ""),   
 Mediana= function(x) format(round(x,2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x,2), big.mark =""))

| Cuadro 11. Distribución del ingreso de la ocupación principal según la Intensidad de la ocupación | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Intensidad | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Sobreocupado | 2319 | 20.42 % | 32262.92 | 28000 | 25821.85 | 80.04 | 0.50 |
| Ocupado pleno | 5471 | 48.18 % | 27983.49 | 25000 | 26053.23 | 93.10 | 0.34 |
| No\_trabajo | 2476 | 21.8 % | 25480.10 | 23000 | 19642.93 | 77.09 | 0.38 |
| Subocupado | 1090 | 9.6 % | 12093.22 | 8000 | 13335.47 | 110.27 | 0.92 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

tabla\_relacion <- datos\_f %>%   
 group\_by(Relacion) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_relacion <- tabla\_relacion %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Relacion, N, Freq, everything())  
  
  
flextable((tabla\_relacion),   
 col\_keys = names(tabla\_relacion)) %>%   
 add\_header\_lines(.,values = c("Cuadro 12. Distribución del ingreso de la ocupación principal según la Relación con el Jefe/a del hogar"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x,2), big.mark = ""),   
 Mediana= function(x) format(round(x,2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x,2), big.mark =""))

| Cuadro 12. Distribución del ingreso de la ocupación principal según la Relación con el Jefe/a del hogar | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Relacion | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Jefe/a | 5773 | 50.84 % | 30897.04 | 28000 | 27516.83 | 89.06 | 0.32 |
| Conyuge | 2643 | 23.27 % | 26977.53 | 24000 | 22832.30 | 84.63 | 0.39 |
| No\_fliar | 42 | 0.37 % | 23083.33 | 17000 | 18249.60 | 79.06 | 1.00 |
| Madre/Padre | 45 | 0.4 % | 22779.11 | 21600 | 16065.90 | 70.53 | 0.22 |
| Otros | 318 | 2.8 % | 20411.01 | 18000 | 14935.16 | 73.17 | 0.48 |
| Hermano/a | 166 | 1.46 % | 18356.02 | 15500 | 14680.53 | 79.98 | 0.58 |
| Hijo/a | 2259 | 19.89 % | 18230.45 | 15000 | 15186.47 | 83.30 | 0.64 |
| Nieto/a | 110 | 0.97 % | 16368.82 | 13450 | 14608.83 | 89.25 | 0.60 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

tabla\_tipo <- datos\_f %>%   
 group\_by(Tipo) %>%   
 summarise(N= n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_tipo <- tabla\_tipo %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Tipo, N, Freq, everything())  
  
  
flextable((tabla\_tipo),   
 col\_keys = names(tabla\_tipo)) %>%   
 add\_header\_lines(.,values = c("Cuadro 13. Distribución del ingreso de la ocupación principal según el Tipo de establecimiento"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%   
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x,2), big.mark = ""),   
 Mediana= function(x) format(round(x,2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x,2), big.mark =""))

| Cuadro 13. Distribución del ingreso de la ocupación principal según el Tipo de establecimiento | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tipo | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Estatal | 3210 | 28.27 % | 36229.29 | 30000 | 24951.87 | 68.87 | 0.75 |
| Otro | 127 | 1.12 % | 28046.17 | 28000 | 17880.01 | 63.75 | 0.01 |
| Privado | 8019 | 70.61 % | 22986.39 | 18000 | 23113.78 | 100.55 | 0.65 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

A partir del Gráfico 4 y Cuadro 6 se observa que más de la mitad de las personas relevadas son hombres (55,64%) y que en promedio los mismos tienden a percibir un ingreso mayor por su ocupación principal que las mujeres ($29372,35 y $23543,33 respectivamente). A su vez, los ingresos de los hombres presentan una mayor variabilidad que el de las mujeres, aspecto que se registra al observar el CV de cada uno.

Con respecto al nivel educativo, aproximadamente el 43,49% de las personas cuentan con el secundario completo, mientras que un 28,73% con primaria completa y un 24,74% con universitaria completa. Por otro lado, se observa que a mayor nivel educativo las personas tienden, en promedio, a percibir ingresos más elevados (Gráfico 4 y Cuadro 7). Este aspecto también se observa al analizar cómo se distribuye el ingreso en función de la calificación del puesto de trabajo y de la intensidad de la ocupación. Aquellos puestos que requieren una mayor preparación tienden, en promedio, a percibir ingresos más altos. Los profesionales ganan un 62,17% más que los técnicos y estos últimos un 30,41% más que los operarios. A su vez, más de la mitad de las personas (52,04%) se desempeñan como operarios, mientras que sólo el 8,54% lo hace como profesionales (Gráfico 5 y Cuadro 8). En relación a la intensidad, se observa que casi la mitad de las personas (48,18%) son ocupados plenos, es decir que trabajaron más de 35 y menos de 45 horas semanales o bien menos de 35 horas pero que no desearon trabajar más, mientras que el 20,12% son sobreocupadas (trabajaron más de 45 horas semanales), el 9,6% subocupadas (trabajaron menos de 35 horas semanales pero desearon trabajar más), y el restante 21,8% se trata de aquellos ocupados que no trabajaron en la semana. En relación a los ingresos, se observa que las ocupaciones de mayor intensidad (sobreocupado y ocupado pleno) son las que perciben ingresos más elevados (Gráfico 6 y Cuadro 11).

Por otra parte, el 76,07% de las personas relevadas se desempeña como obrera o empleada, mientras que aproximadamente el 3% lo hace como patrón y el 20,95% restante como cuentapropista. Estos últimos son los que en promedio perciben menores ingresos, seguidos por los empleados y luego por los patrones (Gráfico 6 y Cuadro 10). Además, el 70,61% se desempeña en el sector privado mientras que el 28,27% lo hace en el público, siendo en este último donde, en promedio, se perciben los ingresos más elevados (Gráfico 7 y Cuadro 13).

Por último, se observa que el 50,84% de las personas se perciben como jefe/as del hogar y, en promedio, son los que registran mayores ingresos, seguidos por los conyuges quienes representan el 23,23% y ganan un 14,52% menos que los jefe/as (Gráfico 7 y Cuadro 12). Con respecto al estado conyugal, las personas casadas son las que perciben mayores ingresos, seguidas por aquellas unidas y solteras (Gráfico 5 y Cuadro 9).

## Análisis de la variable Categoría

list\_cno <- datos\_f %>%  
 dplyr::select(Categoria) %>%   
 unique()  
  
#Creación de variable de grandes grupos de caracter de ocupación según CNO   
#https://www.indec.gob.ar/ftp/cuadros/menusuperior/clasificadores/definiciones\_conceptuales\_cno.pdf  
  
datos\_f <- datos\_f %>%   
 mutate(Categoria\_agrup = case\_when(Categoria %in% c("Funcionarios del poder ejecutivo nacional, provincial, municipal y/o departamental",   
 "Funcionarios del poder legislativo nacional, provincial, municipal y/o departamental",  
 "Funcionarios del poder judicial, federal, nacional, provincial, municipal y/o departamental",   
 "Directivos de pequeñas y microempresas", "Directivos de medianas empresas privadas productoras de bienes y servicios", "Directivos de grandes empresas privadas productoras de bienes y servicios",  
 "Directivos de instituciones sociales",   
 "Directivos de organismos, empresas e instituciones estatales") ~ "ocupaciones de dirección",   
  
Categoria %in% c("Ocupaciones de la gestión administrativa, planificación y control de gestión", "Ocupaciones de la gestión jurídico-legal") ~ "ocupaciones de gestión administrativa, de planificación, control de gestión y jurídico-legal",   
  
Categoria %in% c("Ocupaciones de la gestión presupuestaria, contable y financiera") ~ "ocupaciones de gestión presupuestaria, contable y financiera",   
  
Categoria %in% c("Ocupaciones de la comercialización directa", "Ocupaciones de la comercialización indirecta", "Ocupaciones del almacenaje de insumos, materias primas, mercaderías e instrumentos" , "Ocupaciones del transporte", "Ocupaciones del corretaje comercial, venta domiciliaria, viajantes y promotores", "Ocupaciones de la comercialización ambulante y callejera", "Ocupaciones de las telecomunicaciones") ~ "ocupaciones de comercialización,transporte,  
telecomunicaciones y almacenaje",   
  
Categoria %in% c("Ocupaciones de la salud y sanidad", "Ocupaciones de la educación", "Ocupaciones de la investigación", "Ocupaciones de la asesoria y la consultaría", "Ocupaciones de la prevención de siniestros (naturales, humanos, productivos) atención del medio ambiente y ecología", "Ocupaciones de la comunicación de masas", "Ocupaciones de los servicios de vigilancia y seguridad civil", "Ocupaciones de servicios policiales", "Ocupaciones de las FFAA Gendarmería y Prefectura", "Ocupaciones de los servicios sociales, comunales, políticos, gremiales y religiosos") ~ "ocupaciones servicios sociales básicos",   
  
Categoria %in% c("Ocupaciones del arte", "Ocupaciones del deporte", "Ocupaciones de servicios de recreación", "Ocupaciones de servicios gastronómicos", "Ocupaciones de los servicios domésticos", "Ocupaciones de los servicios de alojamiento y turismo", "Ocupaciones de servicios de limpieza no domésticos", "Ocupaciones del cuidado y la atención de personas", "Ocupaciones de los servicios sociales varios") ~ "ocupaciones de servicios varios",   
  
Categoria %in% c("Ocupaciones de la producción pesquera", "Ocupaciones de la producción ganadera", "Ocupaciones de la producción forestal", "Ocupaciones de la producción agrícola") ~ "ocupaciones agropecuarias, forestales, de la pesca y de la caza",   
  
Categoria %in% c("Ocupaciones de la producción extractiva", "Ocupaciones de la producción de energía, agua y gas", "Ocupaciones de la construcción edilicia, de obras de infraestructura y de redes de distribución de energía, agua potable, gas, telefonía y petróleo") ~ "ocupaciones de la producción extractiva, energética, de construcción e infraestructura",   
  
Categoria %in% c("Ocupaciones de la producción industrial y artesanal", "Ocupaciones de la producción de software", "Ocupaciones de la reparación de bienes de consumo") ~ " ocupaciones de la producción artesanal, industrial y de reparación de bienes de consumo",  
  
Categoria %in% c("Ocupaciones de la instalación y mantenimiento de maquinaria, equipos y sistemas de la producción de bienes", "Ocupaciones del desarrollo tecnológico productivo", "Ocupaciones de la instalación y mantenimiento de maquinaria, equipos y sistemas") ~ "ocupaciones auxiliares de la producción de bienes y la prestación de servicios",  
TRUE ~ "Ns.Nc"))

#Chequeo de la variable creada  
  
table(datos\_f$Categoria\_agrup)

##   
## ocupaciones de la producción artesanal, industrial y de reparación de bienes de consumo   
## 1180   
## Ns.Nc   
## 66   
## ocupaciones agropecuarias, forestales, de la pesca y de la caza   
## 91   
## ocupaciones auxiliares de la producción de bienes y la prestación de servicios   
## 140   
## ocupaciones de comercialización,transporte,\ntelecomunicaciones y almacenaje   
## 2312   
## ocupaciones de dirección   
## 488   
## ocupaciones de gestión administrativa, de planificación, control de gestión y jurídico-legal   
## 1430   
## ocupaciones de gestión presupuestaria, contable y financiera   
## 475   
## ocupaciones de la producción extractiva, energética, de construcción e infraestructura   
## 1037   
## ocupaciones de servicios varios   
## 2021   
## ocupaciones servicios sociales básicos   
## 2116

#Los casos Ns.Nc son los mismos que no reportaron en la variable Categoría. Lo chequeamos abajo.  
  
datos\_f %>%   
 filter(Categoria\_agrup == "Ns.Nc") %>%   
 dplyr::select(Categoria, Categoria\_agrup)

## # A tibble: 66 x 2  
## Categoria Categoria\_agrup  
## <chr> <chr>   
## 1 Ns.Nc Ns.Nc   
## 2 Ns.Nc Ns.Nc   
## 3 Ns.Nc Ns.Nc   
## 4 Ns.Nc Ns.Nc   
## 5 Ns.Nc Ns.Nc   
## 6 Ns.Nc Ns.Nc   
## 7 Ns.Nc Ns.Nc   
## 8 Ns.Nc Ns.Nc   
## 9 Ns.Nc Ns.Nc   
## 10 Ns.Nc Ns.Nc   
## # ... with 56 more rows

tabla\_categoria <- datos\_f %>%   
 filter(Categoria\_agrup != "Ns.Nc") %>%   
 group\_by(Categoria\_agrup) %>%  
 summarise(N = n(),  
 Media = mean(Ingreso),  
 Mediana = median(Ingreso),  
 Desvío\_std = sd(Ingreso),  
 CV = round(Desvío\_std/Media\*100,2),  
 Asimetría = round(3\*(Media-Mediana)/Desvío\_std,2))  
  
tabla\_categoria <- tabla\_categoria %>%   
 mutate(Freq = paste(round(N/sum(N)\*100,2),"%")) %>%   
 arrange(desc(Media)) %>%   
 dplyr::select(Categoria\_agrup, N, Freq, everything())  
  
  
flextable((tabla\_categoria),   
 col\_keys = names(tabla\_categoria)) %>%   
 add\_header\_lines(.,values = c("Cuadro 18. Distribución del ingreso de la ocupación principal según el caracter ocupacional"))%>%   
 add\_footer\_lines(., "Fuente: EPH 2º Trimestre 2020") %>%  
 set\_table\_properties(., layout = "autofit") %>%   
 set\_formatter(N = function(x) format(round(x, ), big.mark = ""),  
 Media= function(x) format(round(x, 2), big.mark = ""),   
 Mediana= function(x) format(round(x, 2), big.mark = ""),   
 Desvío\_std = function(x) format(round(x, 2), big.mark =""))

| Cuadro 18. Distribución del ingreso de la ocupación principal según el caracter ocupacional | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Categoria\_agrup | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| ocupaciones de dirección | 488 | 4.32 % | 46550.00 | 35000 | 57014.21 | 122.48 | 0.61 |
| ocupaciones auxiliares de la producción de bienes y la prestación de servicios | 140 | 1.24 % | 38624.29 | 35000 | 23238.53 | 60.17 | 0.47 |
| ocupaciones de gestión administrativa, de planificación, control de gestión y jurídico-legal | 1430 | 12.67 % | 37407.90 | 32500 | 24550.36 | 65.63 | 0.60 |
| ocupaciones de gestión presupuestaria, contable y financiera | 475 | 4.21 % | 36280.45 | 33800 | 22508.08 | 62.04 | 0.33 |
| ocupaciones servicios sociales básicos | 2116 | 18.74 % | 33828.26 | 30000 | 19567.77 | 57.84 | 0.59 |
| ocupaciones de la producción extractiva, energética, de construcción e infraestructura | 1037 | 9.19 % | 22709.35 | 17000 | 23201.20 | 102.17 | 0.74 |
| ocupaciones de comercialización,transporte, telecomunicaciones y almacenaje | 2312 | 20.48 % | 22344.84 | 20000 | 18676.22 | 83.58 | 0.38 |
| ocupaciones de la producción artesanal, industrial y de reparación de bienes de consumo | 1180 | 10.45 % | 21367.88 | 18000 | 18708.42 | 87.55 | 0.54 |
| ocupaciones agropecuarias, forestales, de la pesca y de la caza | 91 | 0.81 % | 18900.00 | 12900 | 14811.32 | 78.37 | 1.22 |
| ocupaciones de servicios varios | 2021 | 17.9 % | 14369.51 | 12000 | 12046.77 | 83.84 | 0.59 |
| Fuente: EPH 2º Trimestre 2020 | | | | | | | |

rm(datos\_conti\_b,datos\_conti\_f, datos\_conti\_r,datos)