Proyecto Docentes

Pablo Dinamarca

09/12/2020

Contents

Chapter 1

Introducción. Objetivos de la Investigación.

Objetivo: Identificar y analizar los factores determinantes del aprovechamiento de las TIC por estudiantes, docentes, directores y padres/tutores de la educación escolar básica y de la educación media de instituciones educativas de gestión oficial, que permita al MEC diseñar y ejecutar programas y proyectos para el uso de estas herramientas en las instituciones educativas y en los hogares como apoyo en el proceso de enseñanza aprendizaje.

Chapter 2

Informe técnico. Descripción del diseño muestral.

2.1 Definición de la población objetivo.

La población objetivo considerada son los estudiantes de Educación Escolar Básica (EEB) primer y segundo ciclo, (EEB) tercer ciclo y Educación Media que se encuentran matriculados en instituciones de gestión oficial, docentes y directores de estas instituciones, así como padres/madres o tutores de los estudiantes.

La encuesta será realizada a estudiantes de tercer ciclo de EEB y EM, a los docentes y directores de las instituciones, y a los padres. En el caso de los estudiantes de 1° y 2° ciclo de EEB considerando su corta edad se aplicará el cuestionario autoadministrado a los padres/madres o tutores en el cual se realizan preguntas específicas sobre el aprovechamiento de las TIC de los estudiantes, así como otras preguntas específicas para padres.

2.2 Diseño y Marco Muestral.

2.2.1 Marco muestral

El marco muestral para la selección de la muestra lo constituye la información estadística proveniente del Registro Única del Estudiante (RUE) del Ministerio de Educación y Ciencias con corte 01/07/2020.

2.2.2 Unidades de Muestreo

- La Unidad Primaria de Muestreo (UPM) es la institución educativa

- La Unidad Secundaria de Muestreo (USM) son los estudiantes y sus respectivos directores, docentes y padres

2.2.3 Criterios de Inclusión y Exclusiones de instituciones educativas

2.2.3.1 Inclusión

Serán incluidas las instituciones que contemplen matrícula de la población objetivo.

2.2.3.2 Exclusión

Se fijan como causas de exclusión:

- Instituciones que no contemplen los niveles de interés.
- Instituciones que imparten educación únicamente a estudiantes con necesidades educativas especiales y que tengan dificultad en completar el cuestionario autoadministrado.

2.2.4 Tipo de muestra

La muestra es del tipo probabilística, estratificada, bietápica e independiente en cada área y nivel de estudio. Es probabilística ya que cada unidad de muestreo en las etapas definidas tiene una probabilidad conocida y distinta de cero de ser seleccionada. Es estratificada ya que se han realizado particiones de la población y su unión constituye el total. Es bietápico puesto que se realizan las selecciones en dos etapas definidas.

2.2.5 Variables de estratificación

Tabla 1: Variables de estratificación

Nivel	Variables explícitas de estratificación	Número de estratos explícitos	Variables de estratificación implícitas
1° y 2° ciclo EEB	Área: 1. Urbana 2. Rural	Área: 2	Departamento geográfico Tamaño de la escuela
3° ciclo de EEB y EM	Área: 1. Urbana 2. Rural	Área: 2	Departamento geográfico Tamaño de la escuela

Chapter 3

Variables Estudiadas

Indicador:

- Identifica la utilidad de los diccionarios de datos en la captura de datos estadísticos.
- Elabora un diccionario de datos a partir de los datos de una encuesta.

Objetivo

- Explicar la utilidad de los diccionarios de datos.
- Listar el conjunto de variables de la encuesta, confeccionar un diccionario de las fuentes de datos analizadas.

3.1 Importancia de los Diccionarios de Datos

Los analistas utilizan los diccionarios de datos por cinco razones importantes:

1. Para manejar los detalles en sistemas grandes.
2. Para comunicar un significado común para todos los elementos del sistema.
3. Para documentar las características del sistema.
4. Para facilitar el análisis de los detalles con la finalidad de evaluar las características y determinar dónde efectuar cambios en el sistema.
5. Localizar errores y omisiones en el sistema.

Manejo de detalles: Los sistemas grandes tienen enormes volúmenes de datos que fluyen por ellos en forma de documentos, reportes e incluso pláticas. De manera similar, se llevan a cabo muchas actividades que utilizan los datos existentes o que generan nuevos detalles.

Comunicación de significados: Los diccionarios de datos proporcionan asistencia para asegurar significados comunes para los elementos y actividades del sistema.

Documentación de las características del sistema: Documentar las características de un sistema es la tercera razón para utilizar los sistemas de diccionario de datos. Las características incluyen partes o componentes así como los aspectos que los distinguen.

Facilidades de análisis: La cuarta razón para hacer uso de los diccionarios de datos es determinar si son necesarias nuevas características o si están en orden los cambios de cualquier tipo.

3.2 Elaboración de un Diccionario

A continuacion se elabora un diccionario de datos en R con los detalles en excel.

```r
library(readxl)
library(tidyverse)
library(hrbrthemes)
```

```r
#~~~~~~~~~~#
# Encuesta #
#~~~~~~~~~~#

# Cargamos los datos de la encuesta y seleccionamos los mas relevantes.

setwd("C:/Users/user/Desktop/Universidad/Muestreo/Docentes - TIC/Docentes")

doce_en <- read_excel("Docentes.xlsx")
doce_en <- doce_en[ , c(13, seq(0,9), 11, 12, seq(14, 17), seq(20, 171))]
doce_names <- colnames(doce_en) # lo utilizaremos para el diccionario

# Codificamos los nombres de las columnas.

for(i in 1:168) {
  colnames(doce_en)[i]= paste("do", i, sep = "")
}

# Reducimos los NA del id
doce_en$do1 <- ifelse(doce_en$do1=="NA", doce_en$do13, doce_en$do1)
```

```r
# Codificamos las respuestas de las variables.

doce_sub <- doce_en

for(i in 1:168) {
  doce_sub[i]  <- ifelse(doce_sub[i]=="No", 0,
                         ifelse(doce_sub[i]=="Sí"|doce_sub[i]=="1- No tengo acceso a int
                                1,
                                ifelse(doce_sub[i]=="2- Menos de 1 hora"|doce_sub[i]=="2
                                       2,
                                       ifelse(doce_sub[i]=="3- De 1 a 3 horas"|doce_sub[
                                              3,
                                              ifelse(doce_sub[i]=="4- De 4 a 6 horas"|do
                                                     4,
                                                     ifelse(doce_sub[i]=="5- De 7 a 9 ho
                                                            5,
                                                            ifelse(doce_sub[i]=="6- Más
                                                                   6, NA)))))))
}

# Agregamos las respuestas que no estan codificadas.

for(i in c(seq(1:15), seq(24,29), 33, 41, 48, 60, 62, 69, 84, 106, 123, 168)) {
  doce_sub[i] = doce_en[i]
}
```

Este diccionario es complemento de “dic - docentes.xlsx”

```
#-------------#
# Diccionario #
#-------------#

# Creamos el diccionario de los datos de la encuesta.

dic <- doce_en %>%
  summary.default %>% as.data.frame %>%
  dplyr::group_by(Var1) %>%
  tidyr::spread(key = Var2, value = Freq) %>%
  select(-Class, -Length) %>%
  mutate(Names = 0,
         R_0= 0,
         R_1= 0,
         R_2= 0,
         R_3= 0,
         R_4= 0,
         R_5= 0,
         R_6= 0,
         R_NA=0)

for(i in 1:168) {
  dic$Names[i] = doce_names[i]
  dic$R_0[i] = sum(doce_sub[i]==0 & !is.na(doce_sub[i]))
  dic$R_1[i] = sum(doce_sub[i]==1 & !is.na(doce_sub[i]))
  dic$R_2[i] = sum(doce_sub[i]==2 & !is.na(doce_sub[i]))
  dic$R_3[i] = sum(doce_sub[i]==3 & !is.na(doce_sub[i]))
  dic$R_4[i] = sum(doce_sub[i]==4 & !is.na(doce_sub[i]))
  dic$R_5[i] = sum(doce_sub[i]==5 & !is.na(doce_sub[i]))
  dic$R_6[i] = sum(doce_sub[i]==6 & !is.na(doce_sub[i]))
  dic$R_NA[i] = sum(is.na(doce_sub[i]))
  dic$Total = rowSums(dic[ , 4:11])
  }
```

```r
# Se saca el nombre debido a su longitud para ilustrar los datos.

select(dic, -Names)[17,]
```

Var1	Mode	R_0	R_1	R_2	R_3	R_4	R_5	R_6	R_NA	Total
do17	character	1800	366	0	0	0	0	0	41	2207

```r
# Verificamos que no existen errores en las variables de respuestas.

sum(dic[-c(seq(1:15),seq(24,29),33,41,48,60,62,69,84,106,123,168),]$Total!=2207)
```

```
## [1] 0
```

```r
#~~~~~~~~~~~~~~~~~~~~~~~~#
# Docentes Seleccionados #
#~~~~~~~~~~~~~~~~~~~~~~~~#

# Repetimos el proceso con los datos de los docentes seleccionados

setwd("C:/Users/user/Desktop/Universidad/Muestreo/Docentes - TIC/Docentes")

RRHH <- read_excel("docentes_seleccionados.xls")
RRHH <- RRHH[, c(28, seq(1, 27))]
RRHH_names <- colnames(RRHH) # lo utilizaremos para el diccionario

# Codificamos los nombres de las columnas.

for(i in 1:28) {
  colnames(RRHH)[i]= paste("do", i+168, sep = "")
}

# Unificamos el nombre para unir las bases mas adelante
colnames(RRHH)[1]= "do1"

# Elaboramos otro diccionario de los datos

dic_RRHH <- RRHH %>%
  summary.default %>% as.data.frame %>%
  dplyr::group_by(Var1) %>%
  tidyr::spread(key = Var2, value = Freq) %>%
  select(-Class, -Length)

for(i in 1:28) {
  dic_RRHH$Names[i] = RRHH_names[i]
}

# Se observa el resultado final.
dic_RRHH[1:5,]
```

Var1	Mode	Names
do1	numeric	documento_persona
do170	numeric	codigo_establecimiento
do171	numeric	codigo_institucion
do172	numeric	codigo_oferta_educativa
do173	character	nombre_institucion

Chapter 4

Análisis y consistencias de los resultados de la encuesta

Indicador:

- Expone los errores no relacionados al muestreo más comunes en la puesta en práctica de un plan de muestreo.
- Identifica la utilidad de la integración de los registros administrativos y los datos provenientes de una encuesta.

Objetivo

- Verificación de casos duplicados.
- Verificación de casos que no corresponden.
- Unión de bases de datos con registros administrativos.
- Explicar que tipo de errores encontramos y como podrían afectar las estimaciones futuras.

```
# Union de las bases de datos.

validation <- merge(x = doce_sub, y = RRHH, by ="do1")
```

```r
# Eliminar duplicados y NA de do1 (id)

sum(table(validation$do1)>1)
```

```
## [1] 69
```

```r
sum(is.na(validation$do1))
```

```
## [1] 10
```

```r
v_2 <- validation %>% group_by(do1) %>% slice(1) %>% filter(!is.na(do1))

sum(table(v_2$do1)>1)
```

```
## [1] 0
```

```r
sum(is.na(v_2$do1))
```

```
## [1] 0
```

Existen casos en que la CI de la persona figura dos veces, esto se puede deber a que dicha persona se inscribió dos veces en la encuesta por lo que, ante muchos duplicados, se creara un sesgo mayor a la hora de realizar el análisis de datos.

Por otro lado, también existen id sin identificación (NA), por lo que la persona respondió, pero no anoto sus datos, esto también podría causar sesgos al análisis de datos.

Chapter 5

Cálculo de la tasa de respuesta

Indicador:

- Identifica el tipo de no respuesta (parcial, o total)

Objetivo

- Estimación de la tasa de respuesta por área y departamento.

No se poseen datos del área y departamento dentro de las bases de datos.

- ¿Qué conclusiones podemos sacar?
- ¿Cómo podría afectar a los resultados de la encuesta tasas elevadas de no respuesta?

```
#-------------------#
# Tasa de respuesta #
#-------------------#

# Poblacion que respondio la encuesta = doce_sub
# Poblacion muestral total = RRHH

nrow(doce_sub)/nrow(RRHH)*100
```

```
## [1] 22.76196
```

Podemos notas que la tasa de respuesta es relativamente baja hasta el último corte, por lo que realizar un análisis con los datos actuales podrían no ser representativos de la población objetivo debido a la baja cantidad de datos disponibles de respuestas.

```r
# Creamos un diccionario para la validacion de datos

val_names <- c(doce_names, RRHH_names[-1])

dic_val <- v_2 %>%
  summary.default %>% as.data.frame %>%
  dplyr::group_by(Var1) %>%
  tidyr::spread(key = Var2, value = Freq) %>%
  select(-Class, -Length) %>%
  mutate(Names = 0,
         R_0= 0,
         R_1= 0,
         R_2= 0,
         R_3= 0,
         R_4= 0,
         R_5= 0,
         R_6= 0,
         R_NA=0,
         Total=0)

for(i in 1:195) {
  dic_val$Names[i] = val_names[i]
  dic_val$R_0[i] = sum(v_2[i]==0 & !is.na(v_2[i]))
  dic_val$R_1[i] = sum(v_2[i]==1 & !is.na(v_2[i]))
  dic_val$R_2[i] = sum(v_2[i]==2 & !is.na(v_2[i]))
  dic_val$R_3[i] = sum(v_2[i]==3 & !is.na(v_2[i]))
  dic_val$R_4[i] = sum(v_2[i]==4 & !is.na(v_2[i]))
  dic_val$R_5[i] = sum(v_2[i]==5 & !is.na(v_2[i]))
  dic_val$R_6[i] = sum(v_2[i]==6 & !is.na(v_2[i]))
  dic_val$R_NA[i] = sum(is.na(v_2[i]))
  dic_val$Total = rowSums(dic_val[ , 4:11])
}
```

```r
# Se saca el nombre debido a su longitud para ilustrar los datos.

select(dic_val, -Names)[17,]
```

Var1	Mode	R_0	R_1	R_2	R_3	R_4	R_5	R_6	R_NA	Total
do17	numeric	1603	333	0	0	0	0	0	29	1965

```r
# Verificamos que no existen errores en las variables de respuestas.

sum(dic_val[-c(seq(1:15),seq(24,29),33,41,48,60,62,69,84,106,123,seq(168,195)),]$Total!=
```

```
## [1] 0
```

Chapter 6

Resultados

1. Resúmenes gráficos de las variables analizadas
2. Estimaciones globales y desagregadas de las distintas variables
3. Comentarios y conclusiones generales

NOTA: Todos los análisis se realizaron a partir de una tasa de respuesta baja, por lo que los gráficos y datos no son representativos del objetivo del trabajo, la finalidad de estos será de ejemplificar un análisis de datos para el periodo final del proyecto.

```
#------------------#
# Datos Relevantes #
#------------------#

table(v_2$do178)
```

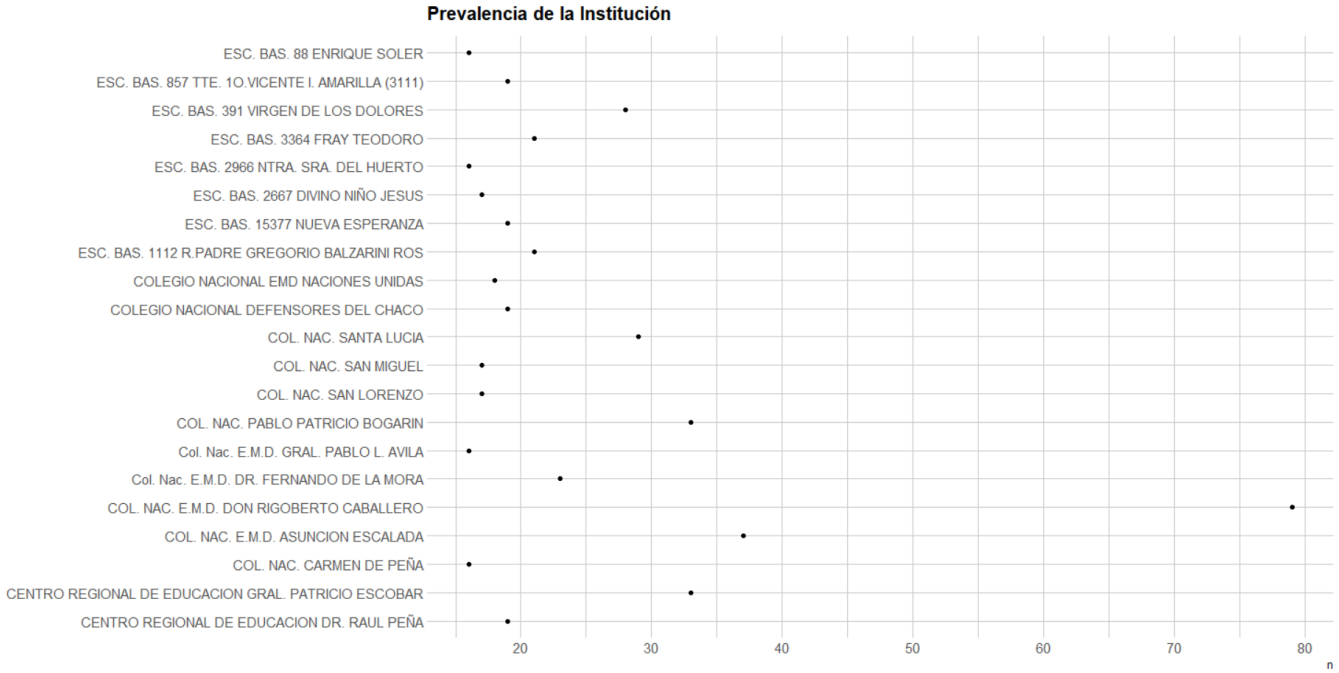
```
##
##  Femenino Masculino
##      1400       565
```

```
mean(v_2$do178=="Femenino")*100
```

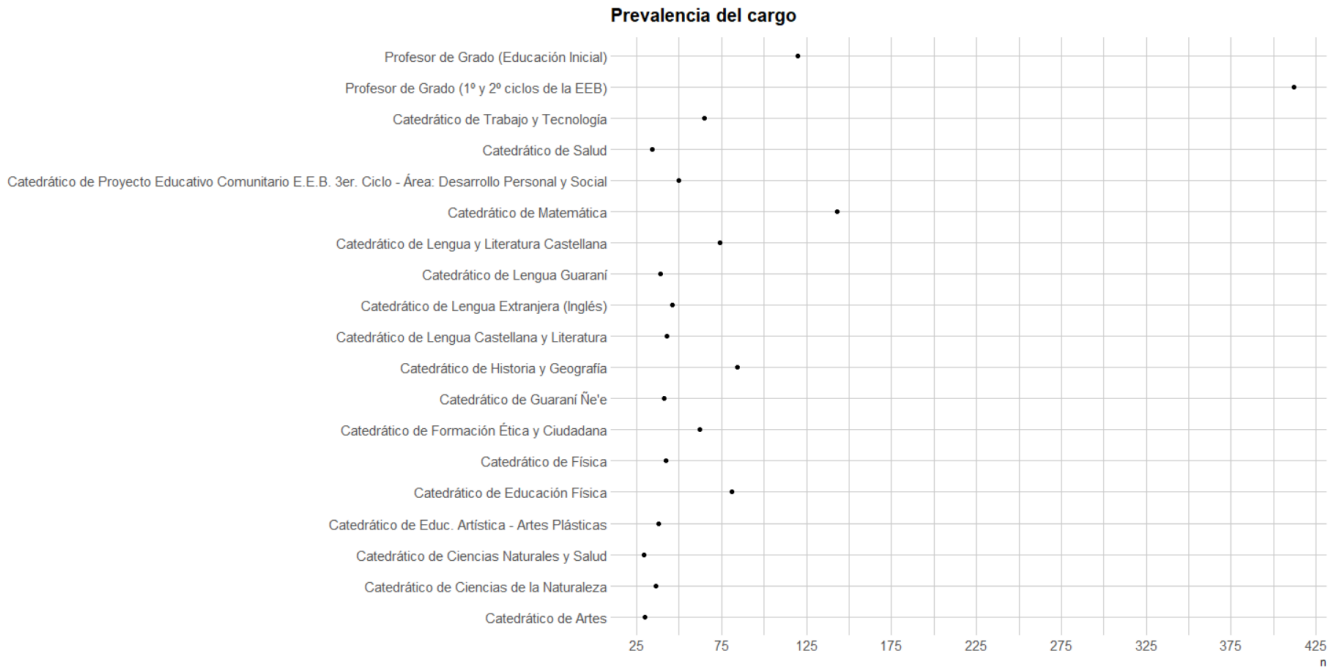
```
## [1] 71.24682
```

```
# Predominancia del genero femenino del 71.25%
```

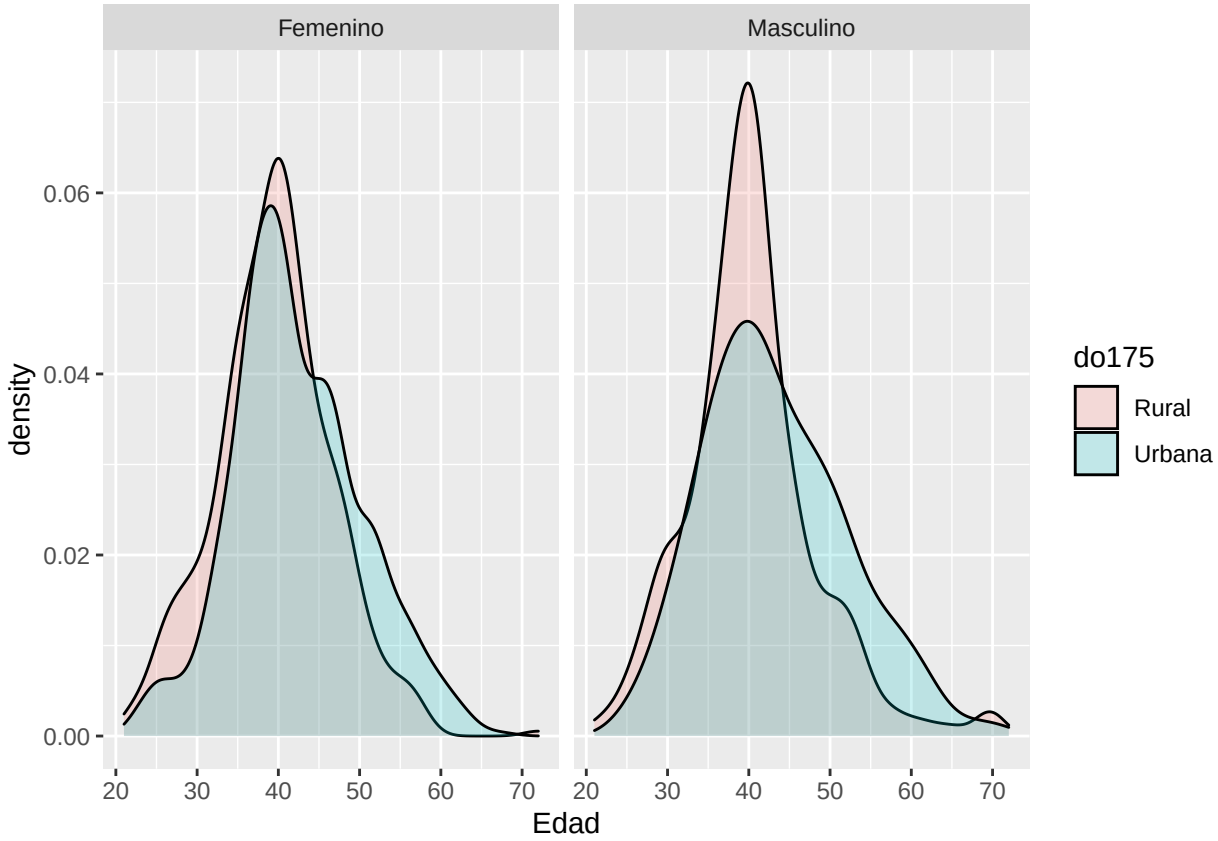
```r
#---------------------#
# Institución y Cargo #
#---------------------#

v_2 %>% select(do173) %>%
  group_by(do173) %>% summarise(n=n()) %>%
  filter(n>15) %>% ggplot(aes(n, do173)) +
  geom_point() + scale_x_continuous(breaks = seq(10, 80, 10)) +
  ggtitle("Prevalencia de la Institución") +
  ylab("") +
  theme_ipsum() +
  theme(plot.title = element_text(size=15))

# El gráfico nos ilustra la predominancia de los docentes por
# institución educativa que se encuentran en los datos de la encuesta.
```



```r
v_2 %>% select(do186) %>%
  group_by(do186) %>% summarise(n=n()) %>%
  filter(n>25) %>% ggplot(aes(n, do186)) +
  geom_point() + scale_x_continuous(breaks = seq(25, 425, 50)) +
  ggtitle("Prevalencia del cargo") +
  ylab("") +
  theme_ipsum() +
  theme(plot.title = element_text(size=15))

# También podemos observar la predominancia del sector educativo en que
# ejercen dichos docentes.
```



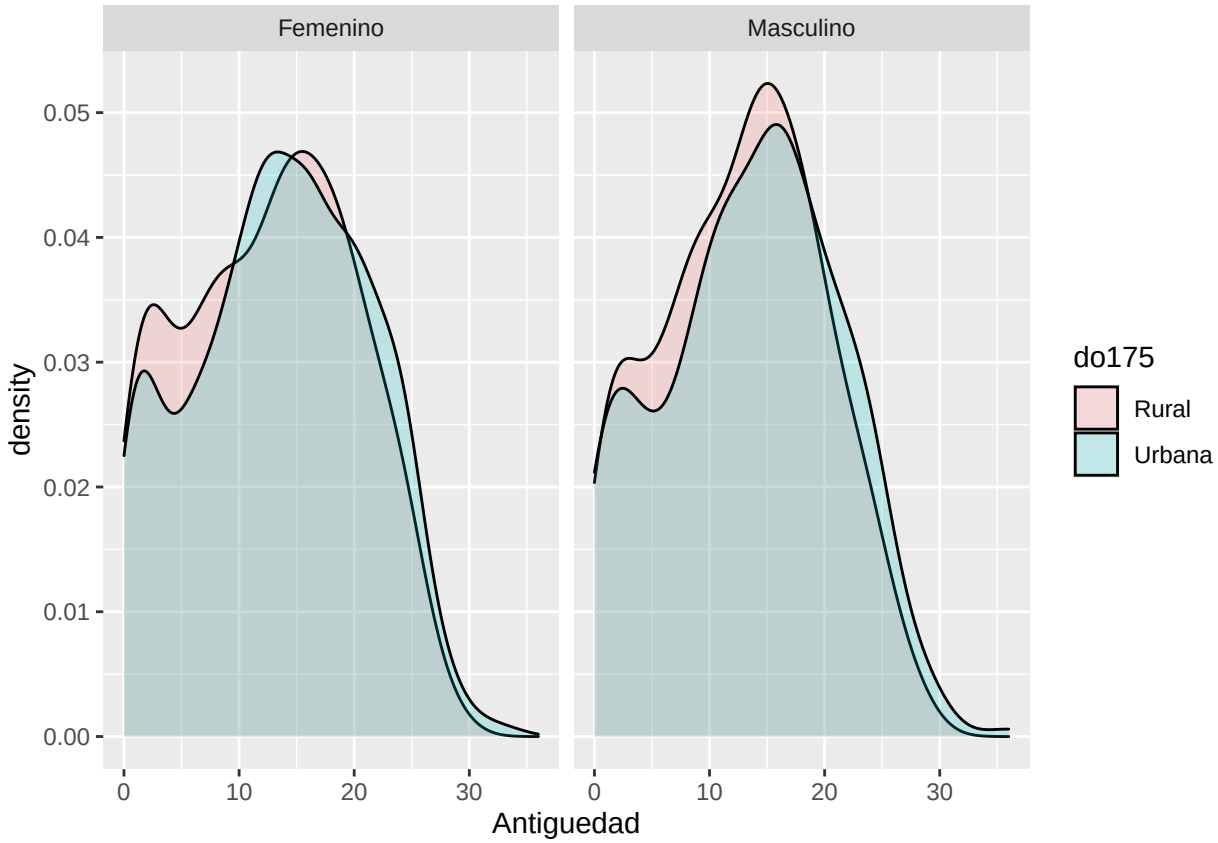
```r
#------------------#
# Edad y Antiguedad #
#------------------#

v_2 %>%
  ggplot(aes(as.numeric(do179), fill=do175)) +
  geom_density(alpha=0.2)  +
  scale_x_continuous(breaks = seq(20, 80, 10)) +
  facet_grid(~do178) +
  xlab("Edad") +
  theme(plot.title = element_text(size=15))
```



El presente grafico nos ilustra la distribución de la edad clasificado por sexo y filtrado por zona.

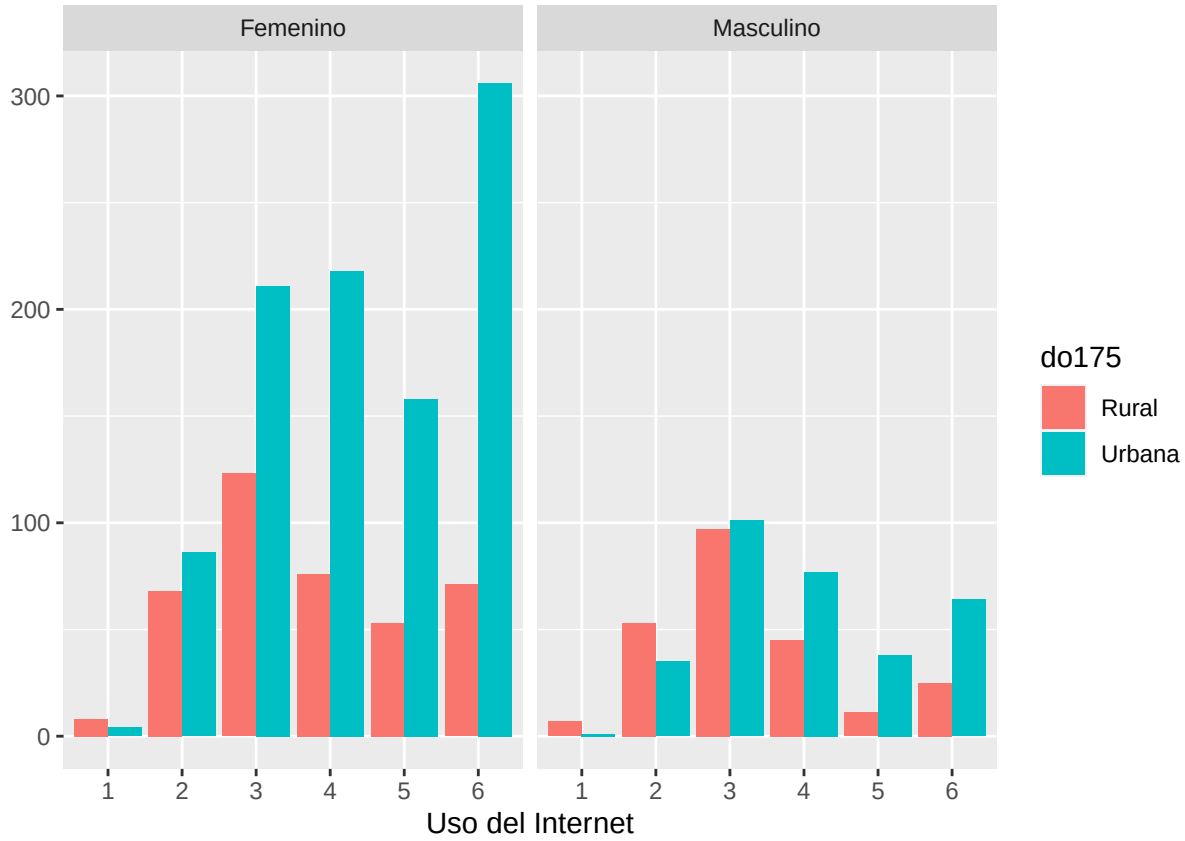
Notamos que el género femenino tiene una distribución de la edad más equilibrada que el masculino, en el cual existe una prevalencia de hombres en el área rural mayor que en la urbana.

```r
v_2 %>%
  ggplot(aes(as.numeric(do180), fill=do175)) +
  geom_density(alpha=0.2)  +
  scale_x_continuous(breaks = seq(0, 80, 10)) +
  facet_grid(~do178) +
  xlab("Antiguedad") +
  theme(plot.title = element_text(size=15))
```



Al comparar los años de antigüedad, notamos que siguen una distribución más equilibrada entre las zonas, pero el recuento del género masculino muestra una mayor antigüedad que el del femenino.

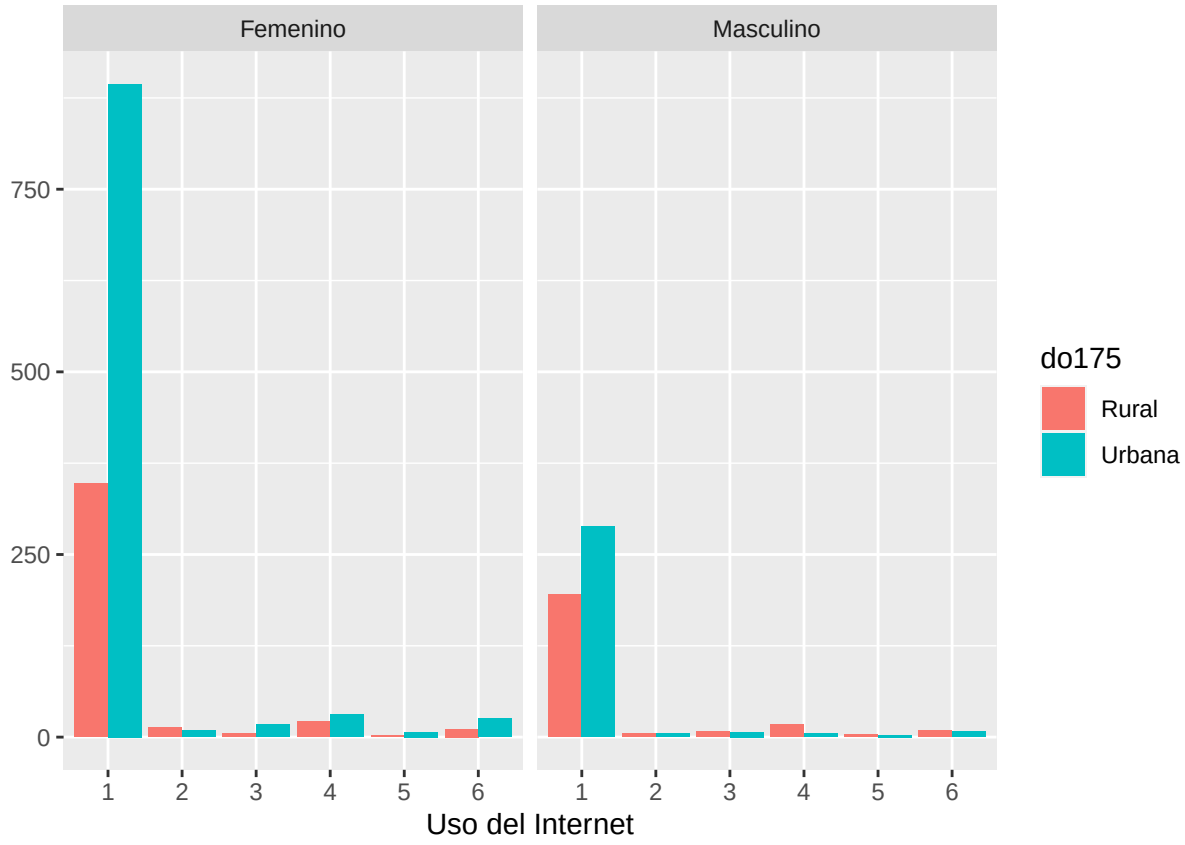
```r
#-----------------#
# Uso de Internet #
#-----------------#

# 1- No tengo acceso a internet
# 2- Menos de 1 hora
# 3- De 1 a 3 horas
# 4- De 4 a 6 horas
# 5- De 7 a 9 horas
# 6- Más de 9 horas

v_2 %>% filter(!is.na(do49)) %>%
  group_by(do49) %>%
  ggplot(aes(as.character(do49), fill=do175)) +
  geom_bar(position="dodge") +
  facet_grid(~do178) +
  xlab("Uso del Internet") +
  ylab("") +
  theme(plot.title = element_text(size=15))
```



El este gráfico se compara el uso de internet clasificado por género y filtrado por zona.

Debido a que los datos están desbalanceados, el género femenino adopta una postura con mayor nivel de horas al día en ambas zonas además de amplificarse más en el área urbana que en la rural.

```r
v_2 %>% filter(!is.na(do61)) %>%
  group_by(do61) %>%
  ggplot(aes(as.character(do61), fill=do175)) +
  geom_bar(position="dodge") +
  facet_grid(~do178) +
  xlab("Uso del Internet") +
  ylab("") +
  theme(plot.title = element_text(size=15))
```




```r
# 1- En su casa
# 2- En la casa de un amigo
# 3- En la casa de un pariente
# 4- En la institución educativa
# 5- En algún lugar público (shopping, café, biblioteca, plaza, etc.)
# Otro
```

Con un énfasis similar al grafico anterior, notamos que la mayoría de los docentes utilizan el internet en su casa.

NOTA: La siguiente tabla tiene como finalidad servir de ejemplo como una forma de agrupación de las variables de la encuesta. Las mismas podemos dividirlas por:

1. Uso de la Tecnología.
2. Uso de las TICs. (Se puede desagregar por institución, motivo etc.)
3. Opiniones. (Calidad de la enseñanza, TIC para el cambio etc.)
4. Habilidades con la tecnología y el internet.
5. Actividades. (Como docente y con portales de uso educativo.)

```
#----------------------#
# Uso de la tecnología #
#----------------------#

# Generamos el recuento de equipos tecnológicos en el hogar

Tecno <- data.frame(N=as.numeric(lapply(17:23, function(i){
  sum(!is.na(v_2[,i]) & v_2[,i]==1)
})))

# Preparamos el primer dataset

Tecno <- Tecno %>%
  mutate(Names=c("PC", "Notebook", "Tablet", "Celular", "TV", "Radio", "No tiene"),
         Porcentaje = round((N/1965*100), digits = 2)) %>%
  select(Names, N, Porcentaje)

# Agregamos el uso frecuente de cada tecnología

Tecno <- Tecno %>% mutate(A_diario=c(as.numeric(t(data.frame(lapply(c(50,51,52,54,56,53
  data.frame(table(v_2[,i])) %>% select(Freq)
}))[,1]), NA),
Una_semanal=c(as.numeric(t(data.frame(lapply(c(50,51,52,54,56,53), function(i){
  data.frame(table(v_2[,i])) %>% select(Freq)
}))[,2]), NA),
Rara_vez=c(as.numeric(t(data.frame(lapply(c(50,51,52,54,56,53), function(i){
  data.frame(table(v_2[,i])) %>% select(Freq)
}))[,3]), NA),
Ninguna_vez=c(as.numeric(t(data.frame(lapply(c(50,51,52,54,56,53), function(i){
  data.frame(table(v_2[,i])) %>% select(Freq)
}))[,4]), NA))
```

```
# Observamos el resultado final
Tecno
```

Names	N	Porcentaje	A_diario	Una_semanal	Rara_vez	Ninguna_vez
PC	333	16.95	506	143	225	1062
Notebook	1148	58.42	1164	215	187	370
Tablet	61	3.10	98	38	134	1666
Celular	1151	58.58	1524	18	45	349
TV	965	49.11	1165	207	319	245
Radio	526	26.77	476	308	455	697
No tiene	17	0.87	NA	NA	NA	NA

De esta tabla se puede analizar la predominancia de los equipos tecnológicos en el hogar además de la necesidad de la población.

Un ejemplo de ello es el uso de la PC, donde 333 personas cuentan con una pc en su hogar pero son 649 los que lo necesitan con más urgencia.

6.1 Conclusiones

A partir del análisis realizado se puede concluir una serie de características de los datos que actualmente no serían relevantes debido a la baja tasa de respuesta. A pesar de ello, este trabajo tiene como finalidad presentar un análisis a modo de ejemplo para luego implementarlo en el trabajo final y sacar conclusiones relevantes.

Chapter 7

Bibliografía y fuentes de datos empleadas

1. Ingenieria De Software - DICCIONARIO DE DATOS
2. METADATA CUESTIONARIO PARA DOCENTES