Tarea 1

Erick Zarza

18 de agosto de 2017

Control 1

Estudio sobre el prestigio de la Ocupación en Canadá e Índice Socioeconómico de Duncan.

1. Análisis exploratorio de Datos

El conjunto de datos tiene los siguientes faltantes y observaciones duplicadas, respectivamente:

```
## OCUPACION EDUCACION   INGRESO   MUJERES PRESTIGIO     CENSO      TIPO
##         0         0         0         0         0         0         4
```

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

Hay 4 faltantes y no ha duplicados, sin embargo hay 2 profesiones con la misma clave del censo

```
## [1]  7 72
```

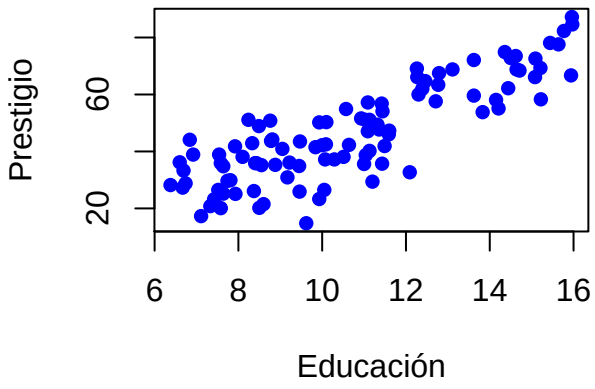
```
##         OCUPACION EDUCACION INGRESO MUJERES PRESTIGIO CENSO TIPO
## 6      PHYSICISTS     15.64   11030    5.13      77.6  2113 prof
## 7      BIOLOGISTS     15.09    8258   25.65      72.6  2113 prof
## 71 SLAUGHTERERS.1      7.64    5134   17.26      25.2  8215   bc
## 72 SLAUGHTERERS.2      7.64    5134   17.26      34.8  8215   bc
```

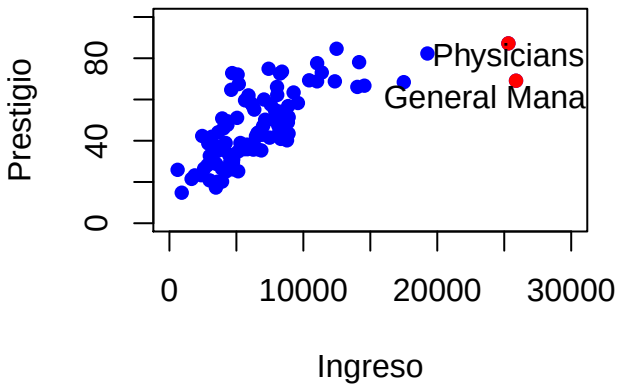
El número total de faltantes es 4 de las 102 observaciones, por lo tanto la métrica de completitud indica que faltan el 3.92% de los datos y todos son de la variable "Tipo", además que no hay duplicados.

Es de llamar la atención que el primer estudio se realizó con ingresos mayores de $3500 y en los datos presentados hay 17 observaciones con un ingreso promedio menor a $3500 :

```
##                     OCUPACION INGRESO
## 28              NURSING.AIDES    3485
## 36                    TYPISTS    3148
## 38           TELLERS.CASHIERS    2448
## 41                FILE.CLERKS    3016
## 42              RECEPTIONISTS    2901
## 45       TELEPHONE.OPERATORS    3161
## 52               SALES.CLERKS    2594
## 53                    NEWBOYS     918
## 54 SERVICE.STATION.ATTENDANT    2370
## 60                      COOKS    3116
## 63                BABYSITTERS     611
## 64                LAUNDERERS    3000
## 65                   JANITORS    3472
## 68              FARM.WORKERS    1656
## 73                    CANNERS    1890
## 75         TEXTILE.LABOURERS    3485
## 84     SEWING.MACH.OPERATORS    2847
```

Detección visual de outliers

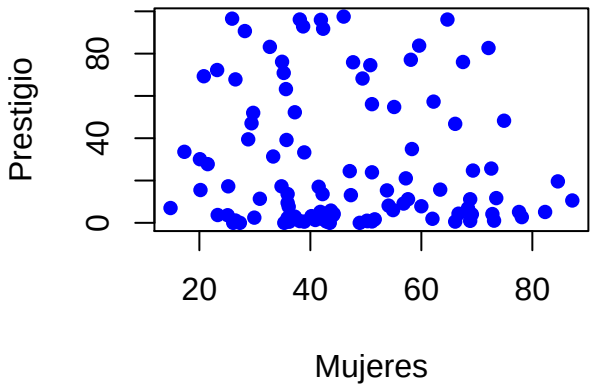


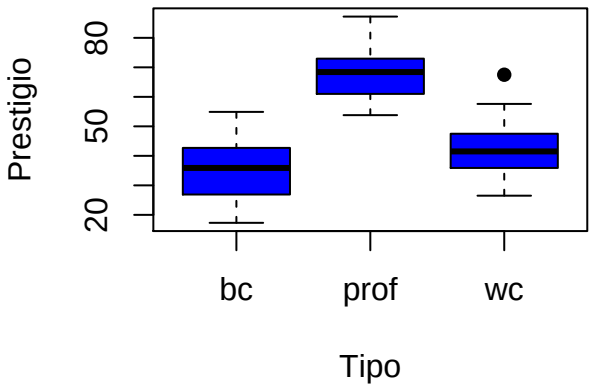


Las observaciones a las que les corresponde el punto rojo (probable outlier) en el comparativo de entre el Prestigio y el Ingreso son:

```
##            OCUPACION EDUCACION INGRESO MUJERES PRESTIGIO CENSO TIPO
## 2  GENERAL.MANAGERS      12.26   25879    4.02      69.1  1130 prof
## 24       PHYSICIANS      15.96   25308   10.56      87.2  3111 prof
```

```
##              OCUPACION EDUCACION INGRESO MUJERES PRESTIGIO CENSO TIPO
## 31 MEDICAL.TECHNICIANS     12.79    5180   76.04      67.5  3156   wc
```



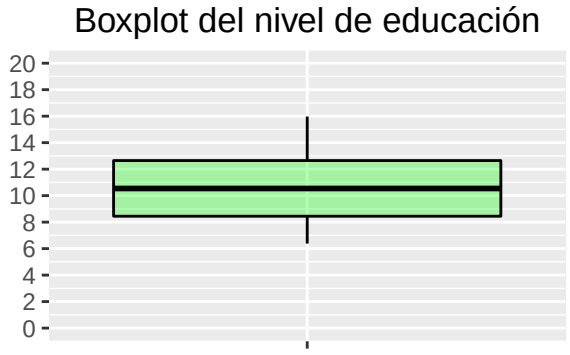


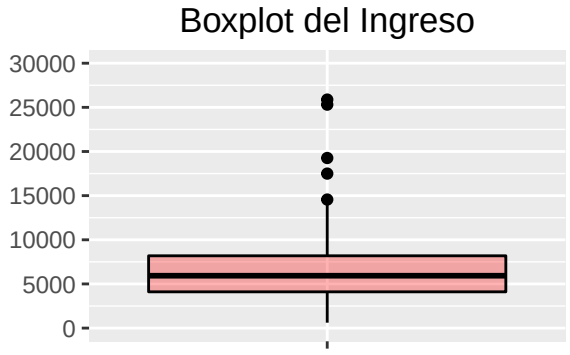
El posible outlier de "WC" es Medical Technicians.

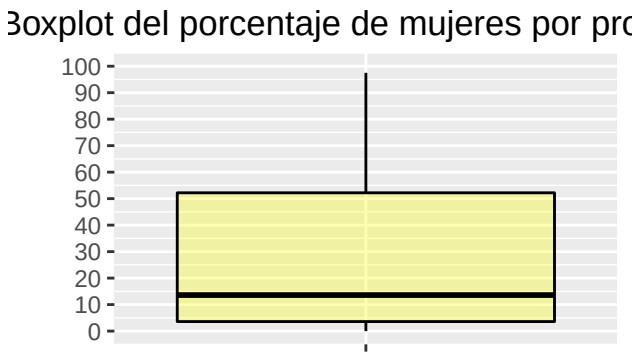
El valor mínimo, primer cuartil, mediana, media 3er cuartil y máximo de cada variable es:

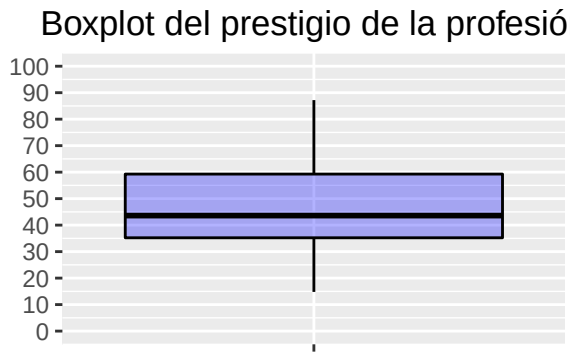
```
##    EDUCACION         INGRESO         MUJERES         PRESTIGIO
##  Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80
##  1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23
##  Median :10.540   Median : 5930   Median :13.600   Median :43.60
```

```
##  Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83
##  3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27
##  Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20
```









2. Clasificación de variables

```
## 'data.frame':    102 obs. of  7 variables:
##  $ OCUPACION: Factor w/ 102 levels "ACCOUNTANTS",..: 41 40 1 70 18 62 11 4 19 53 ...
##  $ EDUCACION: num  13.1 12.3 12.8 11.4 14.6 ...
##  $ INGRESO  : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
##  $ MUJERES  : num  11.16 4.02 15.7 9.11 11.68 ...
##  $ PRESTIGIO: num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ CENSO    : int  1113 1130 1171 1175 2111 2113 2113 2141 2143 2153 ...
##  $ TIPO     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

Los datos contienen 7 variables distintas: Ocupación, educación, ingreso, mujeres, prestigio, censo, tipo. De las cuales las variables ocupación, censo y tipo son variables cualitativas; así como educación, mujeres y prestigio son del tipo cuantitativas continuas, e ingreso una variable cuantitativa discreta en este caso.

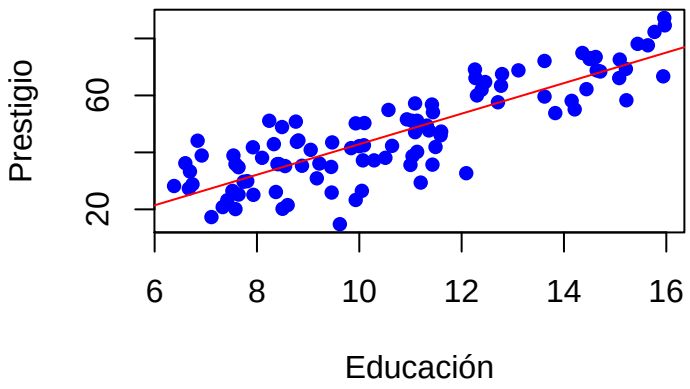
```
##             OCUPACION EDUCACION INGRESO MUJERES PRESTIGIO CENSO TIPO
## 1  GOV.ADMINISTRATORS     13.11   12351   11.16      68.8  1113 prof
## 2    GENERAL.MANAGERS     12.26   25879    4.02      69.1  1130 prof
## 3         ACCOUNTANTS     12.77    9271   15.70      63.4  1171 prof
## 4 PURCHASING.OFFICERS     11.42    8865    9.11      56.8  1175 prof
## 5            CHEMISTS     14.62    8403   11.68      73.5  2111 prof
```

```
##        OCUPACION EDUCACION INGRESO MUJERES PRESTIGIO CENSO TIPO
```

```
## 98   BUS.DRIVERS      7.58    5562    9.47      35.9  9171   bc
## 99  TAXI.DRIVERS      7.93    4224    3.59      25.1  9173   bc
## 100 LONGSHOREMEN      8.37    4753    0.00      26.1  9313   bc
## 101  TYPESETTERS     10.00    6462   13.58      42.2  9511   bc
## 102  BOOKBINDERS      8.55    3617   70.87      35.2  9517   bc
```

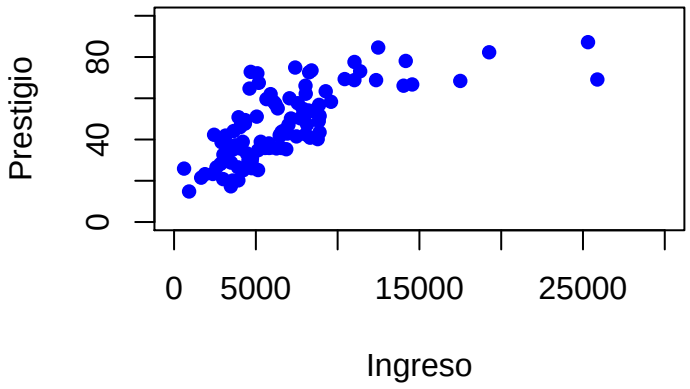
3. Explicación del prestigio respecto al nivel educativo

```
##
## Call:
## lm(formula = datbr$PRESTIGIO ~ datbr$EDUCACION)
##
## Coefficients:
##     (Intercept)  datbr$EDUCACION
##         -10.732            5.361
```

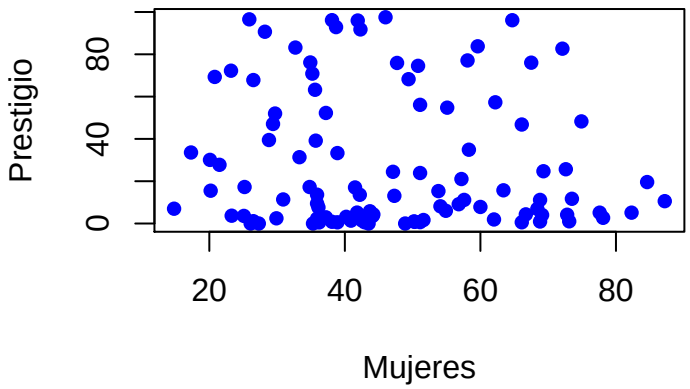



```
##
## Call:
## lm(formula = datbr$PRESTIGIO ~ datbr$EDUCACION)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -10.732      3.677  -2.919  0.00434 **
## datbr$EDUCACION    5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72
## F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

Se puede observar que los resultados de graficar el prestigio y la educación por profesión, nos muestran la relación que existe entre ambas, puesto que mayor educación implica mayor prestigio. Considerando el modelo lineal de R, se rechaza la H_0 en favor de la H_1 para β_0 y β_1.



El prestigio parece que se puede explicar a través del ingreso. Ya que mientras mayor es el ingreso, tambien es mayor el prestigio.



La concentración de los datos muestra una mayor acumulación en un prestigio bajo, pero no parece posible encontrar una tendencia.