PARCIAL FINAL
CORTE 1

Angie Caterine Sarmiento

1. Construya una tabla que resuma,cada variable númerica mínimo, Q_1, mediana, media, Q_3, máximo y desviación estándar, y para la variable cualitativa las frecuencias y los porcentajes de cada una de sus categorías.

   Variables númericas.

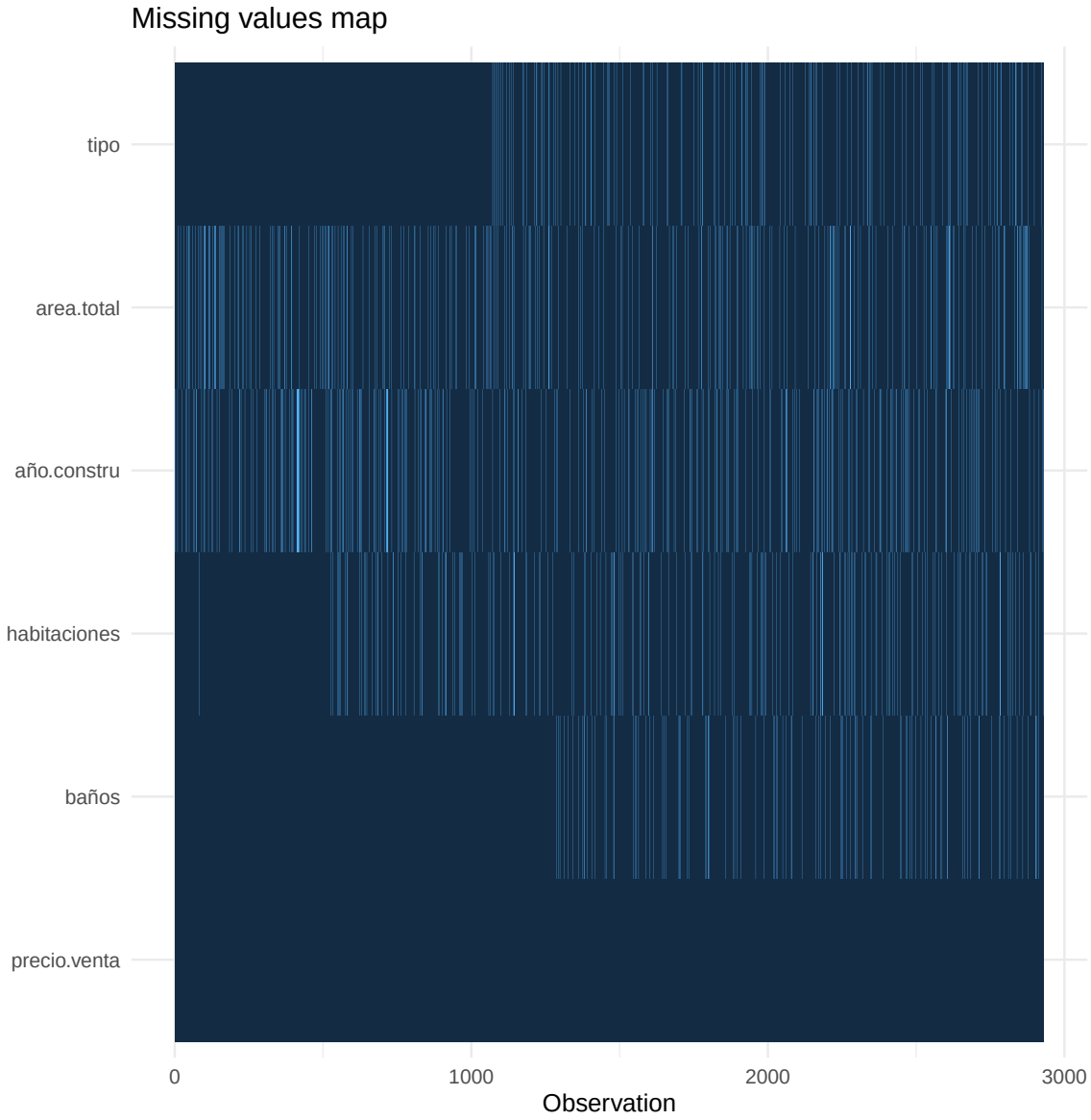
```
##            area.total año.construu habitaciones baños precio.venta
## Minimo        1300.00      231.00         1.00  1.00     12789.00
## Q.25%         7500.00     1954.00         1.00  1.00    129900.00
## Mediana       9532.00     1974.00         1.00  1.00    161900.00
## Media        10120.42     1970.83         1.37  1.15    181989.68
## Q.75%        11609.50     2001.00         1.00  1.00    215000.00
## Maxímo      215245.00     2025.00         4.00  3.00    745000.00
## Desviacion    7859.90       48.95         0.76  0.43     80359.01
```

   Variable categórica.

```
## # A tibble: 4 x 3
##   tipo        Total Porcentaje
##   <fct>       <int>      <dbl>
## 1 apartamento  1228      41.9
## 2 casa         1463      49.9
## 3 Duplex        101       3.45
## 4 <NA>          138       4.71
```

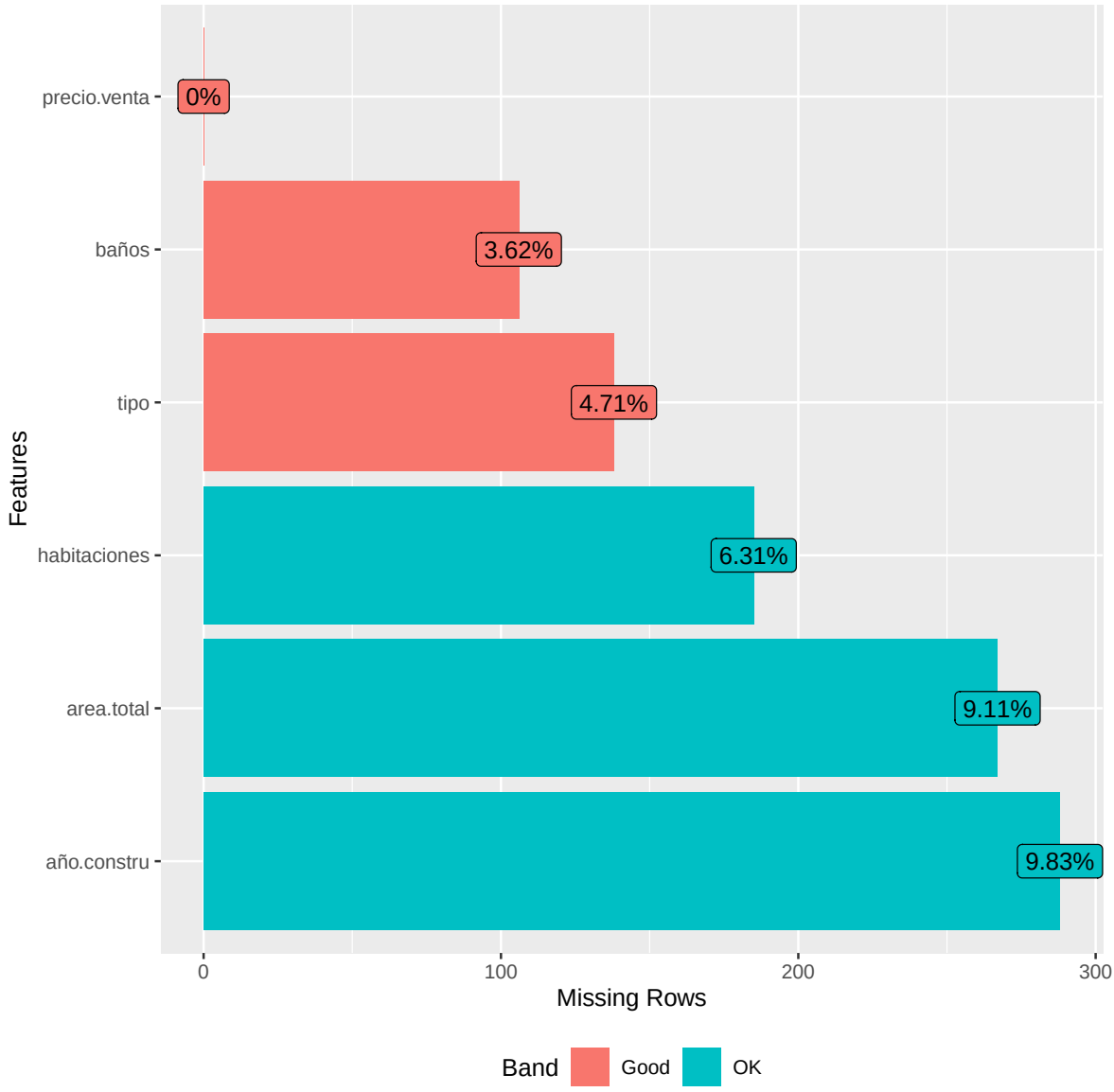
2. Determine el número y el porcentaje de valores faltantes de todo el conjunto de datos. Presente a través de un gráfico apropiado la distribución de este tipo de valores dentro de dicho conjunto.

```
##      Número de NA's Porcentaje de NA's
## [1,]            984            5.59727
```



3. Determine el número y porcentaje de valores faltantes, pero esta vez, dentro de cada una de las seis variables. Construya una tabla de resumén de esta información y un gráfico adecuado.

```
##              Cantidad de NA'S Porcentaje de NA's
## tipo                      138               4.71
## area.total                267               9.11
## año.constru               288               9.83
## habitaciones              185               6.31
## baños                     106               3.62
## precio.venta                0               0.00
```

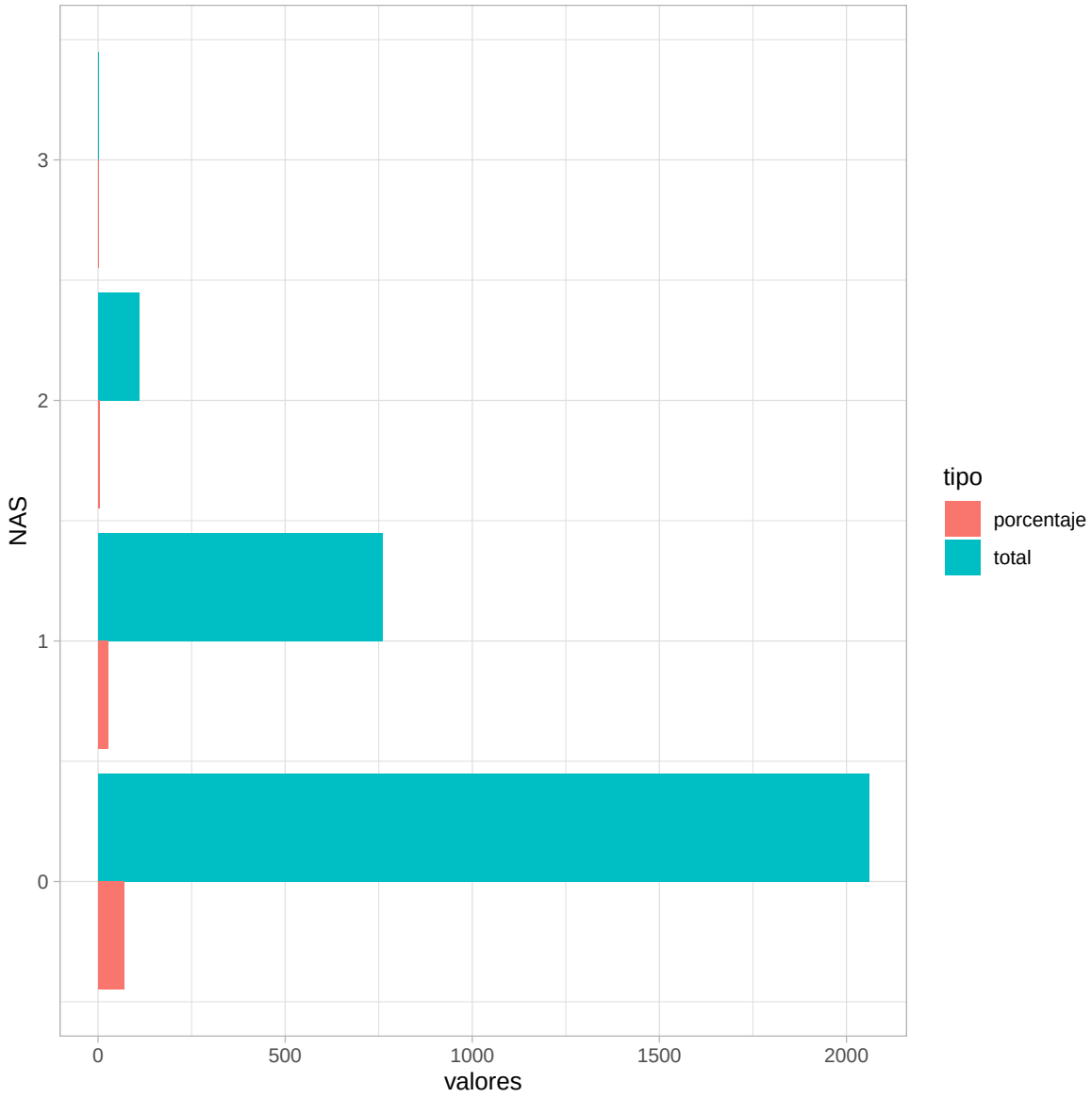


4. Determine el número y el porcentaje de observaciones (fila) con al menos una variable faltante.

```
##   Observaciones.NA.s Porcentaje.NA.s
## 1                111        3.788396
```

5. Determine el número y el porcentaje de observaciones con una, dos, tres, cuatro, cinco o seis valores faltantes (por separado como si fueran categorías). Resuma esta información en un gráfico.

```
## # A tibble: 4 x 3
##     NAS total porcentaje
##   <int> <int>      <dbl>
## 1     0  2059    70.3
## 2     1   760    25.9
## 3     2   109     3.72
## 4     3     2     0.0683
```



6. De acuerdo a los resultados obtenidos en las tareas anteriores, tome una decisión: eliminar filas, columnas (ambas) con NA's o aplicar un método de imputación. Justifique su elección.

   Al revisar los porentajes de NA's en el conjunto de datos se decide realizar imputación.Se encuentra un gran volumen de NA's, si se eliminará se perderia información en la predicción.

7. Si la decisión en la trea 6 fue imputar valores, realice esta tarea:

   a) Cree (y presente) una función que permita realizar imputación simple por muestreo aleatorio sin importar el tipo de variable. Aplique esta función al conjunto de datos con NA's.

```
rand.imput <-function(x){
        missing <- (is.na(x)) #vector booleano
        n.missing <- sum(missing)#Numero de NA's
        x.obs <- x[!missing]#Datos no NA
```

```
        imputed <- x
        imputed[missing] <- sample(x.obs,n.missing,replace = T)
        #Se extrae una muestra aleatoria conocida y se remplazan estos en los NA
        return(imputed)}
ventas_casa1<-data.frame(tipo=rand.imput(ventas_casas$tipo),
                area.total=rand.imput(ventas_casas$area.total),
                año.constru=rand.imput(ventas_casas$año.constru),
                habitaciones=rand.imput(ventas_casas$habitaciones),
                baño=rand.imput(ventas_casas$baños))
head(ventas_casa1)
```

```
##          tipo area.total año.constru habitaciones baño
## 1 apartamento      31770        1960            1    1
## 2 apartamento      11622        1961            1    1
## 3 apartamento      14267        2002            1    1
## 4 apartamento      11160        1968            1    1
## 5 apartamento      13830        1997            1    1
## 6 apartamento       9978        1998            1    1
```

```
tail(ventas_casa1)
```

```
##      tipo area.total año.constru habitaciones baño
## 2925 casa      20000        1960            1    1
## 2926 casa       7937        1984            1    1
## 2927 casa       8885        1914            1    1
## 2928 casa      10441        1999            1    1
## 2929 casa      10010        1974            1    1
## 2930 casa       9627        1993            2    1
```

b) Impute los NA's de todas las variables mediante distribuciones no condicionadas (Algoritmo KNN) y,

```
imputacionKnn<-VIM::kNN(data=ventas_casas,
                        variable = c("tipo","area.total","año.constru","habitaciones",
                                     "baños"),
                        k=5,numFun=mean,catFun=maxCat)
head(imputacionKnn[1:6])
```

```
##          tipo area.total año.constru habitaciones baños precio.venta
## 1 apartamento      31770        1960            1     1       215000
## 2 apartamento      11622        1961            1     1       105000
## 3 apartamento      14267        1973            1     1       172000
## 4 apartamento      11160        1968            1     1       244000
## 5 apartamento      13830        1997            1     1       189900
## 6 apartamento       9978        1998            1     1       195500
```

```
tail(imputacionKnn[1:6])
```

```
##      tipo area.total año.constru habitaciones baños precio.venta
## 2925 casa      20000      1960.0          1.0     1       131000
## 2926 casa       7937      1984.0          1.0     1       142500
## 2927 casa       8885      1958.4          1.0     1       131000
## 2928 casa      10441      1937.0          1.0     1       132000
## 2929 casa      10010      1974.0          1.0     1       170000
## 2930 casa       9627      1993.0          1.6     1       188000
```

c) Impute los NA's de todas las variables mediante imputación múltiple por el algoritmo MICE.

```
multimp.mice<-mice::mice(ventas_casas,m = 5)
```

```
##
##  iter imp variable
##   1   1  tipo  area.total  año.constru  habitaciones  baños
##   1   2  tipo  area.total  año.constru  habitaciones  baños
##   1   3  tipo  area.total  año.constru  habitaciones  baños
##   1   4  tipo  area.total  año.constru  habitaciones  baños
##   1   5  tipo  area.total  año.constru  habitaciones  baños
##   2   1  tipo  area.total  año.constru  habitaciones  baños
##   2   2  tipo  area.total  año.constru  habitaciones  baños
##   2   3  tipo  area.total  año.constru  habitaciones  baños
##   2   4  tipo  area.total  año.constru  habitaciones  baños
##   2   5  tipo  area.total  año.constru  habitaciones  baños
##   3   1  tipo  area.total  año.constru  habitaciones  baños
##   3   2  tipo  area.total  año.constru  habitaciones  baños
##   3   3  tipo  area.total  año.constru  habitaciones  baños
##   3   4  tipo  area.total  año.constru  habitaciones  baños
##   3   5  tipo  area.total  año.constru  habitaciones  baños
##   4   1  tipo  area.total  año.constru  habitaciones  baños
##   4   2  tipo  area.total  año.constru  habitaciones  baños
##   4   3  tipo  area.total  año.constru  habitaciones  baños
##   4   4  tipo  area.total  año.constru  habitaciones  baños
##   4   5  tipo  area.total  año.constru  habitaciones  baños
##   5   1  tipo  area.total  año.constru  habitaciones  baños
##   5   2  tipo  area.total  año.constru  habitaciones  baños
##   5   3  tipo  area.total  año.constru  habitaciones  baños
##   5   4  tipo  area.total  año.constru  habitaciones  baños
##   5   5  tipo  area.total  año.constru  habitaciones  baños
```

```
imput.mice<-complete(multimp.mice)
head(imput.mice)
```

```
##          tipo area.total año.constru habitaciones baños precio.venta
## 1 apartamento      31770        1960            1     1       215000
## 2 apartamento      11622        1961            1     1       105000
## 3 apartamento      14267        1978            1     1       172000
## 4 apartamento      11160        1968            1     1       244000
## 5 apartamento      13830        1997            1     1       189900
## 6 apartamento       9978        1998            1     1       195500
```

```
tail(imput.mice)
```

```
##      tipo area.total año.constru habitaciones baños precio.venta
## 2925 casa      20000        1960            1     1       131000
## 2926 casa       7937        1984            1     1       142500
## 2927 casa       8885        1965            1     1       131000
## 2928 casa      10441        1970            1     1       132000
## 2929 casa      10010        1974            1     1       170000
## 2930 casa       9627        1993            1     1       188000
```

¿Cuál de las tres imputaciones es la más eficiente y adecuada para usted? Argumente y justifique su elección mostrando evidencia numérica o gráfica

Se utiliza el test kolmogorov smirnov para las variables numericas y un test chi-cuadrado para las variables categoricas y así determinar si la distribución de los datos ha cambiado al realizar las imputaciones.

Sistema de hipotesis de la prueba Kolmogorov-smirnov

$H_o : X$ proviene de un modelo probabilistico particular con función de distribución $F(x)$.
$H_a : X$ proviene de cualquier otro modelo probabilistico con función de distribución $G(x) \neq F(X)$.
Matematicamente,este sistema de hipotesis se traduce a:

$$H_o : F(X) = F^*(X)$$

$$H_a : F(X) \neq F^*(X)$$

Sistema de hipotesis de la prueba chi-cuadrado

$$H_o : O_i = E_i$$

$$H_a : O_i \neq E_i)$$

Pruebas de bondad de ajuste para los valores imputados por el muestreo aleatorio simple

```
chisq.test(x=ventas_casas$tipo,y=ventas_casa1$tipo)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ventas_casas$tipo and ventas_casa1$tipo
## X-squared = 5584, df = 4, p-value < 2.2e-16
```

```
ks.test(x=ventas_casas$area.total,y=ventas_casa1$area.total)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$area.total and ventas_casa1$area.total
## D = 0.0050108, p-value = 1
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$año.constru,y=ventas_casa1$año.constru)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$año.constru and ventas_casa1$año.constru
## D = 0.0049582, p-value = 1
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$habitaciones,y=ventas_casa1$habitaciones)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$habitaciones and ventas_casa1$habitaciones
## D = 0.0034981, p-value = 1
## alternative hypothesis: two-sided
```

```
#ks.test(x=ventas_casas£baños,y=ventas_casa1£baños)
```

No se rechaza la hipotesis nuela en las pruebas kolmogorov.

Pruebas de bondad de ajuste para los valores imputados por el metodo KNN

```
chisq.test(x=ventas_casas$tipo,y=imputacionKnn$tipo)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ventas_casas$tipo and imputacionKnn$tipo
## X-squared = 5584, df = 4, p-value < 2.2e-16
```

```
ks.test(x=ventas_casas$area.total,y=imputacionKnn$area.total)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$area.total and imputacionKnn$area.total
## D = 0.009259, p-value = 0.9998
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$año.constru,y=imputacionKnn$año.constru)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$año.constru and imputacionKnn$año.constru
## D = 0.013381, p-value = 0.9647
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$habitaciones,y=imputacionKnn$habitaciones)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$habitaciones and imputacionKnn$habitaciones
## D = 0.023293, p-value = 0.4254
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$baños,y=imputacionKnn$baños)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$baños and imputacionKnn$baños
## D = 0.017249, p-value = 0.7857
## alternative hypothesis: two-sided
```

No se rechaza la hipotesis nula en las pruebas kolmogorov y el valor p es muy cercano a 1,es decir que la distribución de los datos teorica es igual a la impirica que se obtuvo al imputar por el método KNN.

Pruebas de bondad de ajuste para los valores imputados por el metodo MICE

```
chisq.test(x=ventas_casas$tipo,y=imput.mice$tipo)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  ventas_casas$tipo and imput.mice$tipo
## X-squared = 5584, df = 4, p-value < 2.2e-16
```

```
ks.test(x=ventas_casas$area.total,y=imput.mice$area.total)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$area.total and imput.mice$area.total
## D = 0.0044958, p-value = 1
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$año.constru,y=imput.mice$año.constru)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$año.constru and imput.mice$año.constru
## D = 0.0043392, p-value = 1
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$habitaciones,y=imput.mice$habitaciones)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$habitaciones and imput.mice$habitaciones
## D = 0.0024742, p-value = 1
## alternative hypothesis: two-sided
```

```
ks.test(x=ventas_casas$baños,y=imput.mice$baños)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ventas_casas$baños and imput.mice$baños
## D = 0.0042793, p-value = 1
## alternative hypothesis: two-sided
```

No se rechaza la hipotesis nula en las pruebas kolmogorov y el valor p es 1,es decir que la distribución de los datos teorica es igual a la impirica que se obtuvo al imputar por el método MICE.

El mejor metodo de imputación fue imputacion multivariada por ecuaciones encadenadas(MICE),ya que al realizar pruebas de bondad de ajuste entre la distribución teorica de los datos para cada variable y la distribución empirica(la imputada por el metodo MICE) se evidencia que las distribuciónes de las variables no difieren y su valor p es 1.

8. Con los datos ya imputados (por el método que usted escogió como el mejor),ajuste un modelo lineal de regresión múltiple tomando como una variable de respuesta (Precio.venta) y como variable explicativa las demás variables.Reporte el modelo ajustado (con coeficiente de regresión estimados).

```
train<-imput.mice%>%mutate(precio.venta=factor(precio.venta))

modelo<-lm(formula = precio.venta~as.factor(tipo)+area.total+año.constru+
                habitaciones+baños,data=train)
round(coef(modelo),2)
```

```
##           (Intercept)   as.factor(tipo)casa as.factor(tipo)Duplex
##             -4367.00                 -3.94               -136.59
##            area.total           año.constru          habitaciones
##                  0.01                  2.41                 -6.09
##                 baños
##                 -9.15
```

9. Prediga el precio de venta de los siguientes inmuebles con base en el modelo ajustado en el item anterior,y sus correspondientes perfiles, y complete esta tabla:

```
new<-data.frame(tipo=as.factor(c("apartamento","apartamento","casa","Duplex")),
area.total=c(12567,45250,100225,8066),año.constru=c(1965,2010,1905,1942),
habitaciones=c(2,2,1,4),baños=c(1,2,1,2))

prediccion<-predict(object = modelo,newdata = new)
cbind(new,prediccion)
```

```
##          tipo area.total año.constru habitaciones baños prediccion
## 1 apartamento      12567        1965            2     1   455.1718
## 2 apartamento      45250        2010            2     2   831.3126
## 3        casa     100225        1905            1     1  1055.1048
## 4      Duplex       8066        1942            4     2   203.6749
```

Referencias

Ramos, D. (2020).*Detección y tratamiento de datosfaltantes: missing data*