Análisis de caso

Sesión 9

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Venta de viviendas en la Región Metropolitana - Mayo 2020

La base de datos `anuncios` contiene información sobre anuncios de casas en venta en la región metropolitana en la primera semana de Mayo 2020. La información que contiene la base de datos es la siguiente:

- `Identificador:` ID de la vivienda
- `Comuna:` Comuna de la vivienda
- `Tipo:` Tipo de vivienda
- `Habitaciones:` Cantidad de habitaciones en la vivienda
- `Banos:` Cantidad de baños en la vivienda
- `Estacionamientos:` Cantidad de estacionamientos disponibles para la vivienda
- `Metros cuadrados:` Cantidad de metros cuadrados del interior de la vivienda
- `Direccion:` Dirección de la vivienda en venta

Venta de viviendas en la Región Metropolitana - Mayo 2020

Una empresa de corredores de viviendas está muy interesada en obtener inferencias de la dinámica del año 2020 en cuanto a ventas de viviendas en las distintas comunas de la región metropolitana. Se le solicita a usted, realizar análisis de la base de datos y obtener inferencias valiosas para la empresa.

Parte A: Análisis previo

Análisis previo

Esta parte corresponde a trabajo previo de los datos para su análisis. Puede corresponder a "limpieza de datos", recodificación de variables, revisión del formato de las variables, determinar si no existen registros duplicados, etcétera. Esta parte no se incluye en el informe o reporte.

Análisis previo

a.1) Cargue los datos en `R`. Analice la estructura de estos. Observe la variable `Estacionamiento`, esta variable está leída en formato `character` ¿por qué `R` la leyó así? Justifique. Realice la modificación correspondiente de modo que la variable se lea correctamente.

a.1)

```
library(rio) #Cargamos la librería rio
anuncios<-import(file.choose()) #La función import carga los datos
                                #file.choose() abre una nueva ventana
                                #para seleccionar el archivo de datos

dim(anuncios) #Vemos la dimensión de la tabla de datos
[1] 827   9

names(anuncios) #Nombre de las variables, ojo con el Identificador
[1] "Identificador"    "Comuna"           "Link"             "Tipo"
[5] "Habitaciones"     "Banos"            "Estacionamientos" "Metros cuadrados"
[9] "Direccion"

str(anuncios$Estacionamientos) #Revisar siempre el formato y verificar que se lean en el formato adecuado
 chr [1:827] "3" "6" "No" "No" "3" "No" "4" "No" "No" "No" "No" "2" "2" "No" "2" ...

table(anuncios$Estacionamientos) #Vemos que existe un valor "No"
  1   2   3   4   5   6   7   8  No
 91 133  57  37  11   8   2  16 472

unique(anuncios$Estacionamientos) #Valores de la variable
[1] "3"  "6"  "No" "4"  "2"  "1"  "5"  "8"  "7"

#Cambiamos todos los registros "No" a cero:
anuncios$Estacionamientos[which(anuncios$Estacionamientos=="No")]<-0

#Le aplicamos formato as.numeric:
anuncios$Estacionamientos<-as.numeric(anuncios$Estacionamientos)

str(anuncios$Estacionamientos) #Ahora se lee en formato numérico! :)
 num [1:827] 3 6 0 0 3 0 4 0 0 0 ...
```

Análisis previo

a.2) Observe la información que contiene cada variable. Por una parte, ¿existe información sensible en la base de datos? (Entiéndase como *información sensible* aquélla que pudiera ser mal utilizada). Comente por qué es importante ser cuidadosos con la información contenida en las bases de datos. Por otra parte, ¿diría usted que existen variables que quizás no son muy útiles para el análisis solicitado? ¿cuál(es)? ¿por qué?

a.2)

La variable *dirección* indica la dirección exacta de la casa en venta, pudiera ser peligroso subir esta información abiertamente a la red.
La protección del dato es primordial para velar por la seguridad de las personas.

Variables que no son muy útiles para el análisis son el link del anuncio, no provee información útil, el tipo de vivienda tampoco pues note que:

```
unique(anuncios$Tipo)
[1] "Casa"
```

Sólo toma el valor "Casa" pues los anuncios corresponden a ventas de casas.

Análisis previo

a.3) Cada observación corresponde a una vivienda única, ¿existen viviendas duplicadas? Comente por qué realizar este análisis es fundamental.

Análisis previo

```
table(table(anuncios$Identificador))

  1
827
```

Se verifica que cada identificador aparece sólo una vez en los datos, por lo tanto, no habrían viviendas duplicadas. Este análisis es importante porque en algunos casos hay registros duplicados los que podrían interferir en el cálculo de las estadísticas, es necesario evaluar dichos registros si no son necesarios para el análisis.

Parte B: Análisis Estadístico

Análisis Estadístico

En esta parte se realizan todos los análisis solicitados, ya nuestros datos se encuentran limpios y trabajables. Esta parte va en el reporte, se trabajan estadísticas, gráficos, tablas, entre otros, y lo más importante es que cada recurso debe estar acompañado de comentarios valiosos e informativos.

Análisis estadístico

b.1) En base a los datos, ¿cuáles son las comunas que registraron con casas en venta en la primera semana de Mayo? ¿en qué comuna(s) se observa(n) mayor y menor cantidad de casas en venta? Comente.

b.1)

```
unique(anuncios$Comuna) #Comunas con anuncios de casas en venta
 [1] "Calera de Tango"     "Cerrillos"           "Cerro Navia"
 [4] "Colina"              "El Bosque"           "El Monte"
 [7] "Huechuraba"          "Independencia"       "La Cisterna"
[10] "La Florida"          "La Granja"           "La Pintana"
[13] "La Reina"            "Lampa"               "Las Condes"
[16] "Lo Barnechea"        "Lo Espejo"           "Lo Prado"
[19] "Macul"               "Padre Hurtado"       "Pedro Aguirre Cerda"
[22] "Providencia"         "Pudahuel"            "Puente Alto"
[25] "Quilicura"           "Quinta Normal"       "Recoleta"
[28] "Renca"               "San Bernardo"        "San Miguel"
[31] "Santiago"            "Vitacura"
```

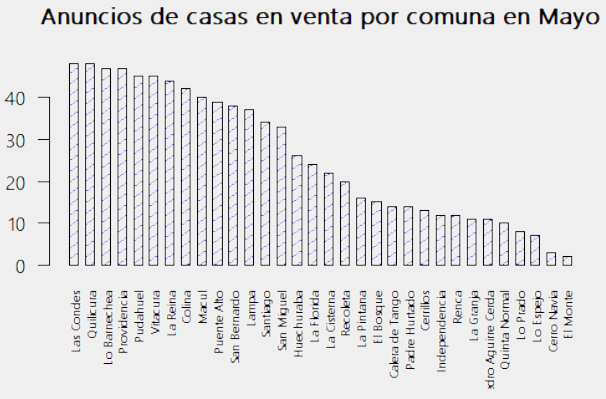
Son 32 comunas en las que se anunciaron casas en venta en la primera semana de Mayo 2020.

b.1)

```
comunas<-data.frame(table(anuncios$Comuna)) #Guardamos la info en una dataframe

comunas<-comunas[order(comunas$Freq, decreasing=TRUE),] #Ordenamos de forma decreciente

head(comunas, 5) #Comunas con mas anuncios de casas en venta
           Var1 Freq
15   Las Condes   48
25    Quilicura   48
16 Lo Barnechea   47
22  Providencia   47
23     Pudahuel   45

tail(comunas, 5) #Comunas con menos anuncios de casas en venta
            Var1 Freq
26 Quinta Normal   10
18       Lo Prado    8
17      Lo Espejo    7
3     Cerro Navia    3
6        El Monte    2

summary(comunas$Freq) #Información de la distribución de anuncios
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   12.00   23.00   25.84   40.50   48.00
```

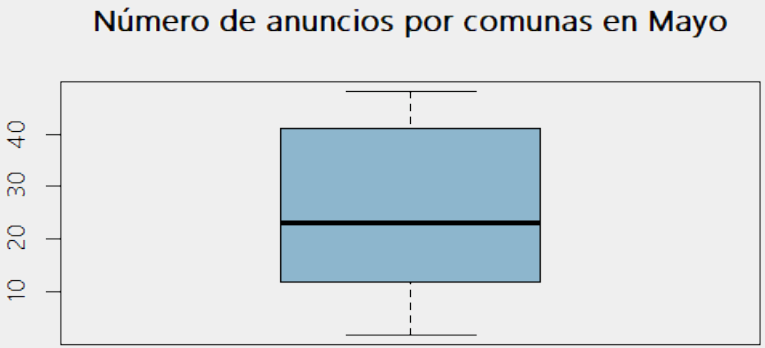
b.1)

```
barplot(comunas$Freq,
        names.arg=comunas$Var1,
        las=2,
        cex.names=0.6,
        col="lightslateblue",
        density=15,
        space=0.8,
        main="Anuncios de casas en venta por comuna en Mayo")
```



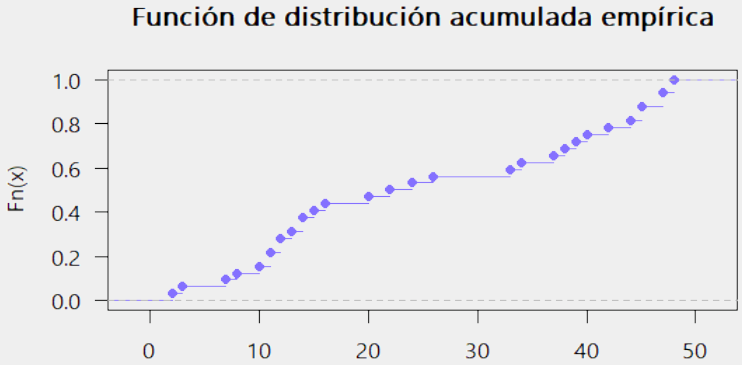
b.1)

```
boxplot(comunas$Freq,
        col="lightskyblue3",
        main="Número de anuncios por comunas en Mayo")

summary(comunas$Freq)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   12.00   23.00   25.84   40.50   48.00
```



b.1)

```
plot(ecdf(comunas$Freq), main="Funcion de distribucion acumulada empirica",

quantile(comunas$Freq, c(0.2, 0.4, 0.6, 0.8))
 20%  40%  60%  80%
11.2 15.4 33.6 43.6
```

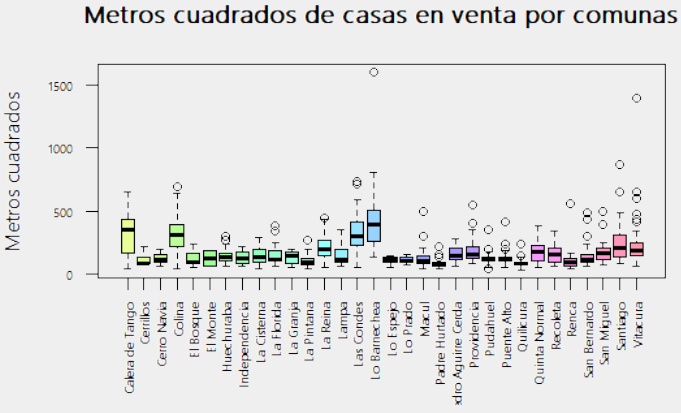


b.2) En términos de metros cuadrados de las viviendas por comuna:

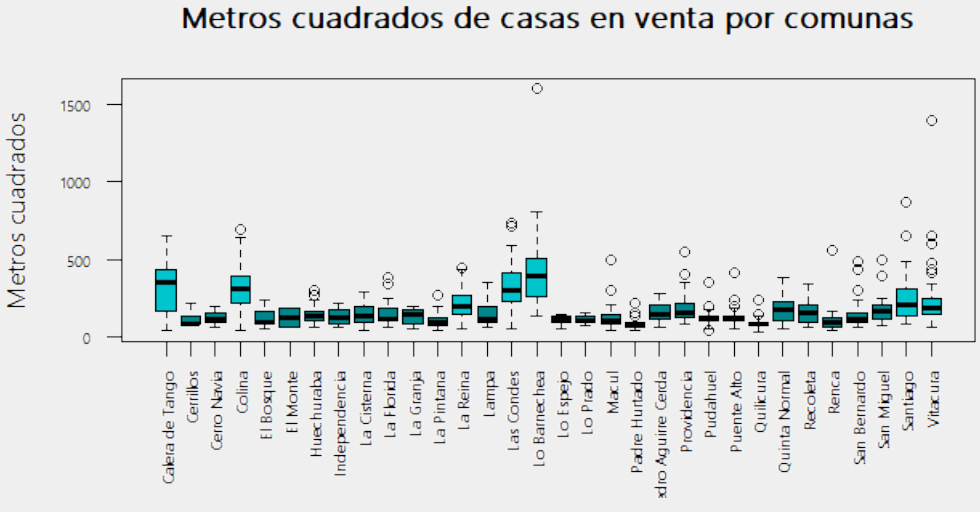
b.2.1) ¿Podría afirmar usted, basado en los datos, que se observan diferencias importantes por comuna? Comente.

b.2.1)

```
boxplot(anuncios$`Metros cuadrados`~anuncios$Comuna,
        xlab="",
        ylab="Metros cuadrados",
        main="Metros cuadrados de casas en venta por comunas",
        las=2,
        cex.axis=0.6,
        col=rainbow(32, alpha=0.4, start=0.2))
```



b.2)

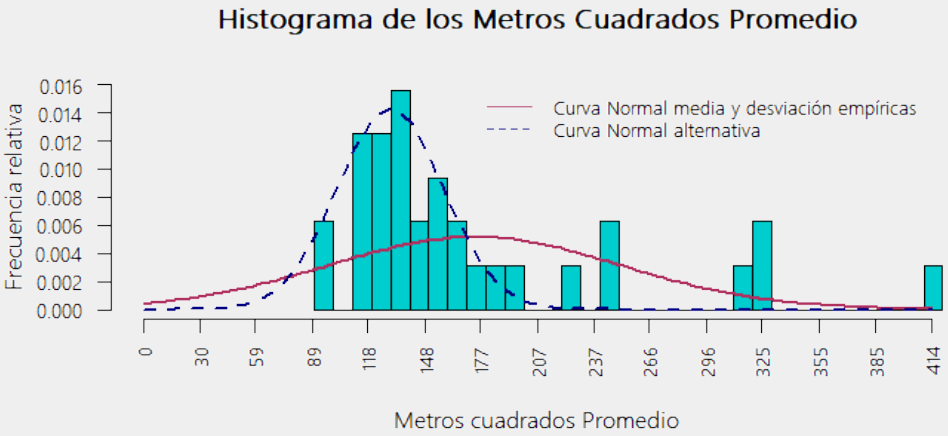

Metros cuadrados de casas en venta por comunas
Metros cuadrados
0
500
1000
1500
Cabrera de Tango
Cerrillos
Cerro Navia
Colina
El Bosque
El Monte
Huechuraba
Independencia
La Cisterna
La Florida
La Granja
La Pintana
La Reina
Lampa
Las Condes
Lo Barnechea
Lo Espejo
Lo Prado
Macul
Padre Hurtado
dro Aguirre Cerda
Providencia
Pudahuel
Puente Alto
Quilicura
Quinta Normal
Recoleta
Renca
San Bernardo
San Miguel
Santiago
Vitacura

b.2.2) Sea Y la variable aleatoria promedio de metros cuadrados por comuna. Realice un gráfico que muestre los valores observados de Y y su frecuencia relativa y sobreponga una curva de una distribución normal. ¿Qué observa?

b.2.2)

```
hist(df$Media, main="Histograma de los Metros Cuadrados Promedio",
     xlab="Metros cuadrados Promedio",
     ylab="Frecuencia relativa",
     las=1,
     breaks=30,
     freq=FALSE,
     col="cyan3",
     xlim=c(0, max(df$Media)),
     ylim=c(0, 0.016),
     axes=FALSE)

axis(1, at=round(seq(0, max(df$Media), len=15),0), cex.axis=0.8, las=2)
axis(2, at=seq(0, 0.016, by=0.002), cex.axis=0.9, las=1)

#Añadimos la curva:
curve(dnorm(x, mean=mean(df$Media), sd=sd(df$Media)),
      col="maroon", lwd=2,add=TRUE, lty=1)

curve(dnorm(x, mean=130, sd=28), #Normal adicional
      col="navyblue", lwd=2,add=TRUE, lty=2)

legend(x=177, y=0.016,legend=c("Curva Normal media y desviación empíricas", "Curva Normal alternativa"),
       col=c("maroon", "navyblue"),cex=0.8,pch="", bty="n", lty=1:2)
```

b.2.2)


Histograma de los Metros Cuadrados Promedio
Frecuencia relativa
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Curva Normal media y desviación empíricas
Curva Normal alternativa
0
30
59
89
118
148
177
207
237
266
296
325
355
385
414
Metros cuadrados Promedio

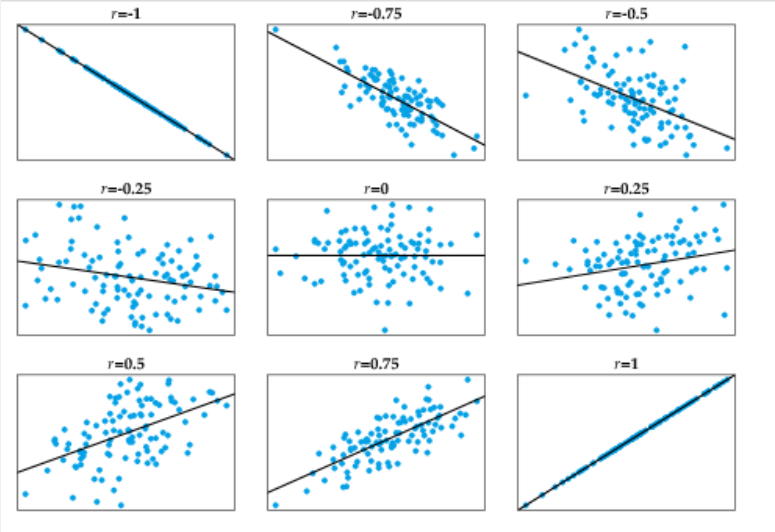
b.3) En términos de metros cuadrados de las viviendas por comuna:

b.3.1) Defina como Z la variable aleatoria promedio de habitaciones por comuna, a través de la base de datos obtenga una muestra de Z. ¿Diría usted que existe correlación entre el promedio de metros cuadrados (Y) y el promedio de habitaciones (Z) por comuna? ¿qué tipo de relación se observa? ¿Tiene sentido? Comente.

b.3.1)

```
df2<-anuncios%>%
  group_by(Comuna) %>%
  summarise(muhabitaciones=mean(Habitaciones),
  mumetros=mean(`Metros cuadrados`))

head(df2)
# A tibble: 6 x 3
  Comuna          muhabitaciones mumetros
  <chr>                    <dbl>    <dbl>
1 Calera de Tango           4.71     324.
2 Cerrillos                 3.85     115.
3 Cerro Navia               3.33     128.
4 Colina                    4.62     317.
5 El Bosque                 3.73     128.
6 El Monte                  4        126.

cor(df2$muhabitaciones,df2$mumetros) #Correlación de Pearson
[1] 0.5665138
```

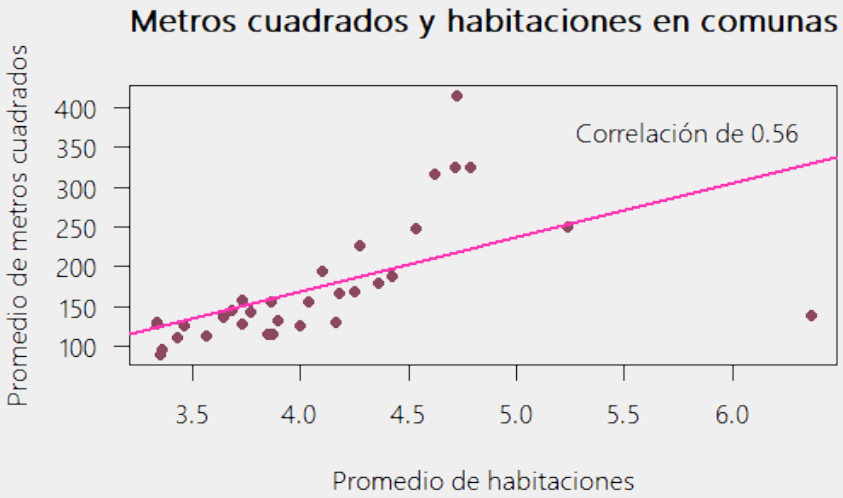
Interpretación Correlación de Pearson



b.3.1)

```
plot(df2$muhabitaciones, df2$mumetros,
     las=1,
     xlab="Promedio de habitaciones",
     ylab="Promedio de metros cuadrados",
     pch=16,
     main="Metros cuadrados y habitaciones en comunas",
     col="palevioletred4")

abline(lm(df2$mumetros~df2$muhabitaciones),
       col="maroon1",
       lwd=2)

text(5.8, 370, "Correlación de 0.56")
```

b.3.1)


Metros cuadrados y habitaciones en comunas
Promedio de metros cuadrados
400
350
300
250
200
150
100
Correlación de 0.56
3.5
4.0
4.5
5.0
5.5
6.0
Promedio de habitaciones

Conclusiones finales

Comentarios objetivos

- Nuestras conclusiones y comentarios deben estar siempre basados en los datos. Nunca hacer juicios de valor u opiniones propias sobre el tema a estudiar, basarnos en lo observable, lo obtenido, lo objetivo.
- Complementar siempre mis conclusiones con gráficos, tablas o estadísticas, pero no utilizar recursos en exceso. Cada recurso debe ser utilizado, por lo tanto utilizar los más importantes, los que ayuden a comprender el fenómeno.
- Los gráficos siempre deben ser claros, los comentarios también y además deben ser precisos, *al grano*.