Practica 2 Limpieza y Validacion de Datos

Carlos A Gutierrez Cruz 10 de junio de 2018

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

PresentaciÃ³n

En esta prÃ¡ctica se elabora un caso prÃ¡ctico orientado a aprender a identificar los datos relevantes para un proyecto analÃtico y usar las herramientas de integraciÃ³n, limpieza, validaciÃ³n y anÃ¡lisis de las mismas.

Competencias

En esta prÃ¡ctica se desarrollan las siguientes competencias del MÃ¡ster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracciÃ³n adecuado a cada situaciÃ³n y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las tÃ©cnicas especÃficas de tratamiento de datos (integraciÃ³n, transformaciÃ³n, limpieza y validaciÃ³n) para su posterior anÃ¡lisis.

Objetivos

Los objetivos concretos de esta prÃ¡ctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resoluciÃ³n de problemas en entornos nuevos o poco conocidos dentro de contextos mÃ¡s amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integraciÃ³n, limpieza y validaciÃ³n) para llevar a cabo un proyecto analÃtico.
- Aprender a analizar los datos adecuadamente para abordar la informaciÃ³n contenida en los datos.
- Identificar la mejor representaciÃ³n de los resultados para aportar conclusiones sobre el problema planteado en el proceso analÃtico.
- Actuar con los principios Ã©ticos y legales relacionados con la manipulaciÃ³n de datos en funciÃ³n del Ã¡mbito de aplicaciÃ³n.
- Desarrollar las habilidades de aprendizaje que les permitan continuar

Desarrollo de la PrÃ¡ctica

1. DescripciÃ³n del dataset. Â¿Por quÃ© es importante y quÃ© pregunta/problema pretende responder?

Para el desarrollo de la prÃ¡ctica se ha escogido un conjunto de datos (dataset) del sitio de Kaggle con la siguiente direcciÃ³n. “https://www.kaggle.com/checoalejandro/autos-consumo-gasolina-mexico/data”

La intenciÃ³n de haber escogido este dataset fuÃ© primero escoger un dataset de algo relacionado con MÃ©xico que es mi paÃs de origen y en segundo porque es conocido que la Ciudad de MÃ©xico es una de las ciudades mÃ¡s pobladas del mundo y que uno de sus principales problemas es la contaminaciÃ³n originados principalmente por los vehÃculos automotores que circulan diariamente en la ciudad.

Bajo el supuesto de que el dataset es una muestra aleatoria y representativa sobre el consumo de gasolina de los vehÃculos automotores de modelos 2011-2018 en MÃ©xico puede ejemplificar perfectamente el uso de tÃ©cnicas con programaciÃ³n en R para el anÃ¡lisis de datos.

El dataset se compone de una muestra de 4617 observaciones y 18 variables que acontinuaciÃ³n se enlistan y describen.

- Marca - Marca frabricante del vehiculo
- Submarca - Sublinea
- Modelo - AÃ±o del vehiculo
- Trans. - Tipo de transmisiÃ³n (manual, automÃ¡tica, CVT)
- Comb. - Tipo de combustible (Diesel, Gasolina)
- Cilindros - NÃºmero de cilindros del vehiculo (4, 6, 8)
- Potencia (HP) - Potencia del motor en caballos de fuerza (HP)
- TamaÃ±o (L) - TamaÃ±o del motor (1.5, 1.8, 2.0)
- CategorÃa - Auto compacto, lujo, deportivo, SUV
- R. Ciudad (km/l) - Rendimiento en Ciudad kilÃ³metros por litro
- R. Carr. (km/l) - Rendimiento en Carretera kilÃ³lmetros por litro
- R. Comb. (km/l) - Rendimiento combinado (ciudad/carretera)
- R. Ajust. (km/l) - Rendimiento ajustado
- CO2(g/km) - Cantidad de emisiÃ³n de Dioxido de Carbono
- NOx (g/1000km) - Cantidad de Oxigeno de Nitrogeno
- CalificaciÃ³n Gas Ef. Inv. - CategorizaciÃ³n de uso de efectivo de combustible
- CalificaciÃ³n Contam. Aire - CategorizaciÃ³n sobre el nivel de contaminaciÃ³n.

Dado la informaciÃ³n anterior es interesante plantearse algunas preguntas como:

Si lo vehiculos cubren las normas de emisiÃ³n planteadas El nivel de emisiÃ³n de los vehiculos tiene relaciÃ³n con el modelo El rendimiento en el consumo de combustible esta relacionado con el modelo o tipo de vehiculo.

Se pueden hacer muchas preguntas al respecto, pero lo importante del ejercicio es mostrar la utilidad que tienen las tÃ©cnicas y herramientas dentro del procesamiento y anÃ¡lisis de los datos.

2. IntegraciÃ³n y selecciÃ³n de los datos de interÃ©s a analizar.

Procederemos a cargar el archivo “Consumo Gasolina Autos Ene 2018.csv” y haremos un anÃ¡lisis exploratorio para observar el tipo de informaciÃ³n contenida que pueda ser de utilidad.

```
dir()
setwd("D:/UOC/02 TipologÃa y ciclo de vida de los datos/Practica 2 Limpieza y Validacion de Datos/Practi
# Carga de datos
data <- read.csv("Consumo Gasolina Autos Ene 2018.csv", header=TRUE)

attach(data)

summary(data)

# Tipo de datos de cada variable
sapply(data, function(x) class(x))

table (data[,6])
```

Dada la informaciÃ³n mostrada, no serÃa Ãºtil considerar los vehiculos con uso de combustible diesel ya que por el numero de observaciones que se tienen no son significantes, asÃ mismo nos centraremos en los autos para lo cual no consideraremos en este trabajo las camionetas o vehiculos utilitarios ya que ademas de estar sujetos a otros niveles ambientales, tienen un rendimiento de combustible mucho menor a los autos a fin de evitar mucha variabilidad y dispersion de los datos durante el anÃ¡lisis.

```
datos2 <- subset(data, Comb.=="Gasolina" & CategorÃa != "CAMIONETAS DE USO MULTIPLE (SUV)")
attach(datos2)
summary(datos2)
```

3. Limpieza de los datos.

3.1. Â¿Los datos contienen ceros o elementos vacÃos? Â¿CÃ³mo gestionarÃas cada uno de estos casos?

las siguientes instrucciones en R nos permitiran conocer si existen ceros o elementos vacios.

```
sapply(datos2, function(x) sum(is.na(x)))
```

Lo cual nos permite apreciar que no existen dentro de las variables de la nuevo dataset(datos2) elementos vacios.

Si hubieran existido elementos vacios dentro de variables se pudieran haber tratado de la siguiente forma.

- Eliminar los registros si fueran pocas las observaciones con elementos vacios
- Completar la informacion con el promedio de las catagorias de la variable si es que no hay mucha variabilidad en los datos.
- Imputacion de valores basado en valores proximos

3.2. IdentificaciÃ³n y tratamiento de valores extremos.

Los siguientes diagramas en R para analizar variables de interes nos ayudan a identificar outliers o valores extremos.

```
#diagramas de boxplot

par(mfrow=c(1,1))

boxplot(CO2.g.km.~ Modelo, data=datos2, main="Emision de CO2 por modelo", xlab="Modelo", ylab="CO2 g/km

boxplot(NOx..g.1000km.~ Modelo, data=datos2,main="Emision de NOx por modelo", xlab="Modelo", ylab="NOx g

boxplot(R..Comb...km.l. ~ Cilindros ,data=datos2,main="Rendimiento Combustible Combinado (Ciudad/Carr.)

boxplot(R..Comb...km.l. ~ Modelo, data=datos2,main="Rendimiento Combustible Combinado (Ciudad/Carr.) po

#grabamos los datos a utilizar para el analisis
write.csv(datos2, "Consumo Gasolina Autos Ene 2018_Nuevo.csv")
```

4. AnÃ¡lisis de los datos.

4.1. SelecciÃ³n de los grupos de datos que se quieren analizar/comparar (planificaciÃ³n de los anÃ¡lisis a aplicar).

Las variables o grupos de datos para analizar y comparar seran:

- Modelo del vehiculo
- Cilindros
- Rendimientos de combustible
- Niveles de emisiÃ³n

```
#Resumen de variables de interes
#AgrupaciÃ³n por Modelo
table(Modelo)
#AgrupaciÃ³n por Num. cilindros
table(datos2$Cilindros)
```

```
#Resumen de medidas de tendencia central
summary(datos2$R..Comb...km.l.)
summary(datos2$CO2.g.km.)
summary(datos2$NOx..g.1000km.)

#Grupos por tipo num de cilindros
datos2.hasta4 <- subset(datos2, Cilindros <= 4)
datos2.mayor4 <- subset(datos2, Cilindros > 4)
```

4.2. ComprobaciÃ³n de la normalidad y homogeneidad de la varianza.

Para la comprobacion de la normalidad existen dos pruebas importantes Kolmogorov-Smirnov y Shapiro-Wilk. Dado que el numero de observaciones es mayor a 50 usaremos la de Kolmogorov-Smirnov

A continuacin se muestra un ejemplo las variables relacionadas con los niveles de emision CO2 y NOx

```
#Prueba de Normalidad

ks.test(CO2.g.km., "pnorm", mean=mean(CO2.g.km., sd=sd(CO2.g.km.)))


ks.test(NOx..g.1000km. , "pnorm", mean=mean(NOx..g.1000km.), sd=sd(NOx..g.1000km.))

#Prueba de F para igualdad de varianzas

var.test(datos2.hasta4$CO2.g.km., datos2.mayor4$CO2.g.km.)
```

El resultado de p-value en la prueba de Kolmogorov-Smirnov es tan pequeÃ±o nos ayuda a determinar que se rechaza la hipÃ³tesis alternativa y se aceptarÃa la hipÃ³tesis nula de que las observacioes provienen de una distrbuciÃ³n normal.

La prueba de F tambien nos ayuda a determinar la igualdad de las varianzas para los grupos de emisiones para automoviles de menos de 4 cilindros con los mayores a 4 cilindros

Lo cual tambiÃ©n puede ejemplificarse de forma grafica

```
par(mfrow=c(2,2))
qqnorm(CO2.g.km., main = "Q-Q Norm Emisiones CO2")
qqline(CO2.g.km., col="red")
hist(CO2.g.km., main = "Histograma Emisiones CO2")
```

4.3. AplicaciÃ³n de pruebas estadÃsticas para comparar los grupos de datos.

En funciÃ³n de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipÃ³tesis, correlaciones, regresiones, etc.

A continuacion vamos a comprobar que los vehiculos de menos de 4 cilindros cumplen con la norma de niveles de emision de Oxido de Nitrogeno NOx. La cual es de 21 g/1000km. Para ello nos apoyaremos de la prueba de t donde las hipotesis a comprobar son H0 La emision de NOx cumple al menos con la norma Ha La emisi

$$ondeNoxesmayoralanorma$$

```
#Prueba de t para una muestra con intervalo de confianza al 95%

t.test(datos2.hasta4$NOx..g.1000km., alternative = "greater", mu=21)

t.test(datos2.mayor4$NOx..g.1000km., alternative = "greater", mu=21)
```

Estadisticamente se puede decir que los vehiculos con mayor a 4 cilindros cumplen con la norma.

Por ultimo realizaremos una relacion si el rendimiento de combustible tiene relacion con el aÃ±o del modelo vehiculo y numero de cilindros. Para ellos revisaremos los coeficientes de correlacion y modelos de regresion

```
cor(datos2$Modelo, datos2$R..Comb...km.l.)

cor(datos2$Cilindros, datos2$R..Comb...km.l.)

pairs(~datos2$Cilindros+datos2$Modelo+datos2$R..Comb...km.l.)

summary(lm(datos2$R..Comb...km.l.~datos2$Modelo+datos2$Cilindros))
```

Conforme al resultado de la funcion de regresion lm(), El rendimiento de combustible de los vehiculos tiene una significante relacion con las varibales modelo y numero de cilindros. Ademas el coeficiente R2 nos muestra que cerca del 67% de los datos quedarian explicados por el modelo.

5. RepresentaciÃ³n de los resultados a partir de tablas y grÃ¡ficas.

```
ryx <-lm(datos2$R..Comb...km.l.~datos2$TamaÃ±o..L.)

linea <- ryx$coefficients[1]+(ryx$coefficients[2]*datos2$TamaÃ±o..L.)
plot(datos2$TamaÃ±o..L.,datos2$R..Comb...km.l.,  xlab = "TamaÃ±o", ylab="Rendimiento km/l")

lines(linea, col=4)
```

La grafica muestra el modelo lineal para la relaciÃ³n que hay entre el rendimiento de comustible dependiendo el tamaÃ±o del motor (1.5litros, 2.0 litros, etc)

6. ResoluciÃ³n del problema. A partir de los resultados obtenidos, Â¿cuÃ¡les son las conclusiones? Â¿Los resultados permiten responder al problema?

El caso prÃ¡ctico nos permitiÃ³ mostrar la aplicaciÃ³n prÃ¡ctica de las tÃ©cnicas y herramientas para el procesamiento y analisis de datos.

Basado en el dataset nos permitio ejemplicar que no existÃa una diferencia en cuanto a los nivele de emisiÃ³n CO2 o Nox independientemente del nÃºmero de cilindros del motor

TambiÃ©n nos permitio ver estadÃsticamente que el rendimientos de combustible de los vehiculos tomados de la muestra tienen una relaciÃ³n significativa con el modelo y numero de cilindros de la maquina.

Por Ãºltimo, tambiÃ©n se puede decir que estadÃsticamente, los vehiculos con numero de cilindros mayor a 4 cumplieron con la norma en el nivel de emisiÃ³n, lo que hace suponer que a pesar de que son maquinas mas grande nos hace suponer que trabajan a mayor estabilidad que los motores con menos cilindros.

Dado lo anterior siempre se puede decir que este tipo de conclusiones estan basadas en la muestra utilizada y para otro tipo de prueba similar siempre es importante llevar acabo una planeaciÃ³ny diseÃ±o para la recolecciÃ³n de los datos sobre el problema a resolver o analizar.