UNIVERSIDAD SANTO TOMÁS
ESTADÍSTICA EXPLORATORIA
TALLER SEGUNDO CORTE. Valor 10%
Fecha de entrega:
Puede ser realizado en parejas

Ejercicio de limpieza y depuración de una base de datos.

Bases de datos: train.csv, y test.csv
DESCRIPCIÓN DE LAS VARIABLES:
The training set contains data we can use to train our model. It has a number of feature columns which contain various descriptive data, as well as a column of the target values we are trying to predict.

The testing set contains all of the same feature columns, but is missing the target value column. Additionally, the testing set usually has fewer observations (rows) than the training set.

las variables que quisiéramos predecir: (Por lo tanto no están en el archivo test)
Casual
Registered
Count

You are provided hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only Information available prior to the rental period.

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, People are able rent.
a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.
In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

This dataset was provided by Hadi Fanaee Tork using data from Capital Bikeshare. We also thank the UCI machine learning repository for hosting the dataset.

If you use the problem in publication, please cite: Fanaee-T, Hadi, and Gama, Joao, Event labeling combining ensemble detectors and background knowledge,
Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

Data Fields
datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday - whether the day is considered a holiday
workingday - whether the day is neither a weekend nor holiday
weather:
1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
Temp: temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated
count - number of total rentals

EJERCICIOS

1. Hay dos cosas qué debe tener en cuenta para poder unir los data frames train1 y test 1? Mencione estas dos cosas.

2. Escriba las instrucciones para solucionar este problema?

3. **I**ntroduzca en la base de datos de Test, nuevas columnas, pues antes de unir las dos bases de datos es necesario que tengan el mismo número de columnas. Por ejemplo para agregar a la base test la columna registered, coloque la instrucción:

4. Es necesario ajustar el nombre de alguna de las variables? O los nombres en las dos bases de datos coinciden? Si es necesario cambiar el nombre de alguna variable,

5. Cree con la función data.frame y con la función colnames, una tabla donde compare los nombres de la base de datos train y test.

6. Una las dos bases de datos

7. Identifique si hay filas repetidas y elimínelas

8. Cree un Data frame donde compare las dimensiones de los data frames originales con la dimensión del nuevo después de unirlas. Use la función dim.

9. **C**uáles variables necesitan ser transformadas en factor?. Como los nombres de los factores son poco ilustrativos, reasigne estos valores con etiquetas más informativas

10. Cuántas variables? ¿cuántas observaciones? con la función dim

11. Instrucción que selecciona los nombres de las variables.

12. Instrucción que me indica el número de filas e instrucción que me indica el número de columnas en un data frame.

13. El promedio del número de bicicletas rentadas (tengan en cuenta que hay datos faltantes)

14. Hacer una tabla con las principales medidas descriptivas para responder la pregunta de si la estación influye o no sobre el número de bicicletas rentadas.

15. Hacer una tabla con las principales medidas descriptivas para responder la pregunta de si el estado del clima influye o no sobre el número de bicicletas rentadas.

16. Instrucción para ver el número de objetos que hay en la memoria en este momento.

17. Vuelva la variable count un factor, con tres categorías, rentasbajas, rentasmedias, rentasaltas,

llamela CountFactor. Es decir, escriba una instrucción que divida en tres partes los datos de la variable count.

18. Agregue esta nueva columna a la base de datos.

19. Elabore una tabla de porcentajes con el paquete base donde cuente cuántos días se registraron ventas Bajas, Medias, Altas

20. Haga el mismo punto anterior con la función correspondiente del paquete janitor utilizando el paquete Janitor (es importante que la variable que categoriza las ventas ya haya sido agregada a la base de datos).

21. Investigue con la ayuda de la función tabyl cómo eliminar la fila de datos faltantes en esta tabla.

22. Tabla cruzada de la variable CountFactor y Season. Con la función table y con la función tabyl del paquete janitor.

23. Seleccione los datos que corresponden a los días festivos

Ejemplo: El promedio del número de bicicletas rentadas en los días festivos.
Observe esta instrucción de ejemplo
mean(data$count[data$holiday=="Holiday"], na.rm = TRUE)
Media_Festivos<-mean(Festivosdata$count, na.rm = TRUE)

24. Haga una tabla donde se visualice el promedio y el Coeficiente de variación para el número de bicicletas rentadas en los días festivos y en los no festivos. Recuerde que hay datos faltantes use el argumento, na.rm = TRUE.

25. En la tabla anterior observa diferencias significativas. Cómo la puede interpretar?

26. Seleccione los datos que se refieren a los días festivo y al mes de verano y calcule un summary.

27. Interprete lo más importante de los resultados obtenidos en el summary del ejercicio anterior.

28. Seleccione los días que corresponde a summer o spring y cuyo número de rentas fueron bajas.

29. Calcule una nueva columna qué indique qué porcentaje de las rentas corresponde a Usuarios registrados y calcule el promedio.

30. Recuerde que casula+registered=count, es decir que el número total de bicicletas rentadas es la suma de las rentas de usuarios registrados y los usuarios casuales.

31. Use la funcíon tabyl para hacer algunas tablas interesantes. Saque porcentajes del total fila, columna o del total general.

32. Hacer un Boxplot del conteo de bicicletas rentadas:

    a) en términos de la estación.
    Coloque colores y un título a este gráfico INTERPRETE

33. A todos los gráficos anteriores agregue una leyenda con la mediana de cada categoría
Observe el siguiente ejemplo:

```
windows() #Abre una ventana gráfica
plot(data$workingday,data$count,
    border = c("blue", "green", "red","pink"),
    main="Diagrama de caja del número de bicicletas \n rentadas en Días laborales y no laborales",
    cex.main=1,    # Tamaño de letra del título)#Elabora el Boxplot
bg="tan1")  # especifica el color del fondo. La lista de los 657 colores disponibles se puede ver con colors())
```

```
levels(data$workingday)#Para visualizar los nombres y escribirlos bien en lo que sigue
tapply(data$count,data$workingday,median,na.rm=T)#Para ver el valor de la mediana por dias lab y no laborales
M1<- median(data$count[data$workingday=="Non workingday"], na.rm = TRUE) #Calcular la mediana del número de biciletas rentadas en la categoría non_Working
M2<- median(data$count[data$workingday=="workingday"], na.rm = TRUE)
la<-paste("Mediana_NonWDay",M1,sep = "\n")#Pegar
lc<-paste("Mediana_WD",M2,sep = "\n")#pegar
text(2,152, lc)# En la ubicación 2, 150 colocar la mediana de los dias nolaborales
text(1,129, la)
text(c(1:nlevels(data$workingday)), labels=paste("n = ",table(data$workingday),sep=""))#Esta es muy chevere para agregar el número de datos en cada caja
```

34. Haga una tabla con el promedio y el CV (coeficiente de variación) del número de bicicletas rentadas.
    a) Ejemplo: En términos de si es festivo o no.

PRIMERA MANERA: "Sin necesidad de usar subset"

```
levels(data$holiday)#Para observar las categorias de la variable y escribirlas bien
mNonHoli<-mean(data$count[data$holiday=="Non Holiday"], na.rm = TRUE)# Para sacar el promedio sin necesidad de hacer el subset
mHoli<-mean(data$count[data$holiday=="Holiday"], na.rm = TRUE)
CVNonHoli<-sd(data$count[data$holiday=="Non Holiday"], na.rm = TRUE)/ mean(data$count[data$holiday=="Non Holiday"], na.rm = TRUE)*100
CVHoli<-sd(data$count[data$holiday=="Holiday"], na.rm = TRUE)/mean(data$count[data$holiday=="Holiday"], na.rm = TRUE)*100
TablaDescriptivas<-data.frame(Media_Festivos=mNonHoli,Media_No_Festivos=mHoli,CV_Festivos=CVHoli,CV_NoFestivos=CVNonHoli)
TablaDescriptivas
```

SEGUNDA MANERA CON LA FUNCIÓN SUBSET

```
NoFestivos<-subset(data,holiday=="Non Holiday") #Selecciona los datos de los días no festivos
Festivos<-subset(data,holiday=="Holiday") #Selecciona los datos de los días festivos
TablaDescriptivas<-data.frame(
            Media_Festivos=mean(Festivos$count, na.rm = TRUE),
            Media_No_Festivos=mean(NoFestivos$count, na.rm = TRUE),
             CV_Festivos=(sd(Festivos$count, na.rm = TRUE)/ mean(Festivos$count, na.rm =
RUE))*100,
           CV_NoFestivos=sd(NoFestivos$count, na.rm = TRUE)/ mean(NoFestivos$count, na.rm = TRUE)*100  )
TablaDescriptivas
```

35. Haga una tabla cruzada donde relacione el número de registros por season y clima y saque porcentaje del total filas.
36. Haga una tabla cruzada donde relacione el número de registros por season y clima y saque porcentaje del total filas.
37. Elabore una tabla de Frecuencias absolutas, relativas, absolutas acumuladas, relativas acumuladas de la varaible count (número total de bicicletas rentadas).

Paso 1. Use la función cut para dividir en tres intervalos Rentas bajas", "Rentas Medias",
Paso2.Agregar una nueva columna con la categoría de las del número de bicicletas rentadas en cada registro (Fecha). Con la función vector cree un vector con el
nombre Categoria_Ventas de longitud apropiada. Una este vector al data frame de datos y llamelo datosnew y finalmente elabore la tabla.

41. Realice un diagrama de dispersión entre humedad y temperatura, acompáñelo de una medida descriptiva apropiada e interprete. ()

42. Cuando no se recomienda usar el operador pipe? https://r4ds.had.co.nz/pipes.html
43. Quién es hadley-wickham? https://priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/

Nota: Por favor interprete muy bien los gráficos y Boxplot obtenidos trate de mencionar las características más importantes.

Recomendados:

https://es.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6th-box-whisker-plots/v/interpreting-box-plots

http://hadley.nz/