1-3-2019

TiCD- PRÀCTICA 2

Descripció : Neteja i validació de dades

Practica 2 – Neteja I validació de dades Carles Maggi -Joan Maggi Gomez

UOC-TCD- Practica 2

Tabla de contenido

**Escribir el título del capítulo (nivel 1)1**

Escribir el título del capítulo (nivel 2)2

Escribir el título del capítulo (nivel 3)3

**Escribir el título del capítulo (nivel 1)4**

Escribir el título del capítulo (nivel 2)5

Escribir el título del capítulo (nivel 3)6

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

Podem trobar els datasets en el següent repositori :

<https://github.com/joanmaggigo/practica2/CSV>

La descripció dels fitxers inicials és la següent :

**HotelsBarcelonaBooking :**

IdHotel : BigInt que identifica de manera unívoca el hotel

Estrelles : Int Número d'estrelles, en cas que n'hi hagi

Nota : Float mitja de l'hotel

Nom : String Nom de l'hotel

Link : String Url de la pàgina de l'hotel

**ComentarisXHotelsBarcelonaBooking :**

IdHotel : BigInt Identifica de manera unívoca el hotel

IdIteració : Int Primera part que identifica un comentari (iteració llista comentaris)

Index : In\* Segon part d'identificació de comentari (número de comentari dins la iteració)

NomHotel : String Nom de l'hotel

Nota : Float Nota que en qualifica el comentari de l'hotel

Comentari Postiu : Text Comentari positiu si n'hi ha

Comentari Negatiu : Text Comentari negatiu si n'hi ha

Data Comentari : Date Data enregistrada del comentari PRAC 1 -Tipologia i Cicle de Dades

**CategoriesXComentariBooking :**

IdHotel : BigInt Identifica de manera unívoca el hotel

IdIteració : Int Primera part que identifica un comentari (iteració llista comentaris)

Index : Int Segon part d'identificació de comentari (número de comentari dins la iteració)

Categoria : Text Categoria amb la que s'ha categoritzat el comentari

# 2. Integració i selecció de les dades d’interès a analitzar.

Un cop definit l’objectiu , caldrà integrà totes les dades en un sol Data set per poder fer totes les probes i testos necessaris.

El data set de treball tindrà les següents atributs:

nomHotel : String, Nom dl Hotel

notapersona : num ,Nota Posada al hotel per la Pêrsona

data: Date, Data del comentari

estrelles: num, Estrelles del Hotel

notaHotel: Nota mitja otorgada al Hotel per Booking

TipusHabitacio: String, Tipus d’Habitació

DuradaEstancia : String, Rang de la estancia

Nits: num, Número de nits d’estància

ProcedenciaComentari: String, Lloc de des de on s’ha fet el comentari

TipusViatge : String, Tipus de Viatge

Acompanyament : String, Tipus d’acompanyament

ViajaConMascota: String, Si viatja en mascota

Season : String, Temporada de l’estada

# 3. Neteja de les dades.

## 3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

## 3.2. Identificació i tractament de valors extrems.

Mirarem els valors extrems de la puntuació donada per un client en referencia a un mateix hotel, no pas tenint en compte totes les puntuacions.

Eliminarem els registres que sobrepassen 3 desviacions estàndards la nota donada per les persones. Ho fem a nivell d'hotel, i no pas a nivell general, perquè entenem que si un hotel és molt dolent potser tothom el puntua amb 1 o 2, que de segur serien outliers si els consideréssim en la totalitat de les puntuacions de tots els hotels, però serien valors força normals a la puntuació pròpia del hotel.

# 4. Anàlisi de les dades.

## 4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

## 4.2. Comprovació de la normalitat i homogeneïtat de la variància.

## 4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l’objectiu de l’estudi, aplicar proves de contrast d’hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d’anàlisi diferents.

# 5. Representació dels resultats a partir de taules i gràfiques.

# 6. Resolució del problema. A partir dels resultats obtinguts, quines son les conclusions? Els resultats permeten respondre al problema?

# 7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s’ha realitzat la neteja, anàlisi i representació de les dades.

Es pot trobar el codi en

<https://github.com/joanmaggigo/practica2>

# TAULA DE CONTRIBUCIONS

|  |  |
| --- | --- |
| **Contribucions** | **Firma** |
| **Investigació Prèvia** | CM,JM (CM->Carles Maggi, JM->Joan Maggi) |
| **Redacció de les Respostes** | CM,JM (CM->Carles Maggi, JM->Joan Maggi) |
| **Desenvolupament del Codi** | CM,JM (CM->Carles Maggi, JM->Joan Maggi) |