Activitat 1: Exploració i preprocés de dades

Enunciat

Estadística Avançada - Semestre 2023.2

Índice

1 Lectura de dades i examinació del tipus de variable **3**
- 1.1 Carregar el fitxer de dades . . . 3
- 1.2 Examinar el tipus de dades . . . 3

2 Normalització de variables qualitatives (text) **3**
- 2.1 Variable raceeth . . . 3

3 Normalització i descripció de variables binàries **3**

4 Normalització de variables quantitatives **3**
- 4.1 Variable readingScore . . . 3
- 4.2 Variable grade . . . 4
- 4.3 Variable schoolSize . . . 4

5 Imputació **4**

6 Mesures característiques de les variables numèriques **4**

7 Arxiu final **4**

Introducció

En aquesta activitat realitzarem l'anàlisi exploratòria i el preprocés del conjunt de dades Pisa. El Programa per a l'Avaluació Internacional d'Estudiants (PISA) és una prova que s'aplica cada tres anys a estudiants de 15 anys de tot el món per avaluar-ne el rendiment en matemàtiques, lectura i ciències. Aquesta prova proporciona una forma quantitativa de comparar el rendiment acadèmic dels estudiants de diferents parts del món.

El conjunt de dades pisa2009.csv conté informació sobre la demografia i les escoles dels estudiants nord-americans que fan l'examen, derivada dels arxius de dades d'ús públic PISA del 2009 distribuïts pel Centre Nacional d'Estadístiques Educatives (NCES) dels Estats Units. Cada fila del conjunt de dades conté la següent informació d'un estudiant:

- grade: El curs que realitza l'estudiant (la majoria dels estudiants de 15 anys als Estats Units són al desè curs).
- male: Si l'estudiant és home (1/0).
- raceeth: La raça/ètnia de l'estudiant.
- preschool: Si l'estudiant va assistir a preescolar (1/0).
- expectBachelors: Si l'estudiant espera fer un grau universitari (1/0).

- motherHS: Si la mare de l'estudiant va completar l'escola secundària (1/0).
- motherBachelors: Si la mare de l'estudiant va obtenir una llicenciatura (1/0).
- motherWork: Si la mare de l'estudiant té feina a temps parcial o complet (1/0).
- fatherHS: Si el pare de l'estudiant va completar l'escola secundària (1/0).
- fatherBachelors: Si el pare de l'estudiant va obtenir una llicenciatura (1/0).
- fatherWork: Si el pare de l'estudiant té feina a temps parcial o complet (1/0).
- selfBornUS: Si l'estudiant va néixer als Estats Units (1/0).
- motherBornUS: Si la mare de l'estudiant va néixer als Estats Units (1/0).
- fatherBornUS: Si el pare de l'estudiant va néixer als Estats Units (1/0).
- englishAtHome: Si l'estudiant parla anglès a casa (1/0).
- computerForSchoolwork: Si l'estudiant té accés a un ordinador per fer les tasques escolars (1/0).
- read30MinsADay: Si l'estudiant llegeix per plaer durant 30 minuts/dia (1/0).
- minutesPerWeekEnglish: El nombre de minuts per setmana que l'estudiant dedica a classe d'anglès.
- studentsInEnglish: El nombre d'estudiants a classe d'anglès d'aquest estudiant a l'escola.
- schoolHasLibrary: Si l'escola d'aquest estudiant té una biblioteca (1/0).
- publicSchool: Si aquest estudiant assisteix a una escola pública (1/0).
- urban: Si l'escola d'aquest estudiant està en una àrea urbana (1/0).
- schoolSize: El nombre d'estudiants a l'escola.
- readingScore: puntuació de lectura de l'estudiant, en una escala de 0 a 1000 punts.

Criteris de verificació i de normalització de les variables:

A continuació es mostren els criteris amb què s'han de netejar les dades del conjunt:

- Reviseu la naturalesa de les variables (text, categòrica, numèrica). En cas que el tipus de variable que ha atorgat R no coincideixi amb el tipus que li correspondria estadísticament, s'haurà de corregir.
- Les variables de caràcter binari han de prendre valors 1 o 0.
- Les variables categòriques amb valors NA no s'alteren. Tampoc no s'eliminen aquestes dades del fitxer per evitar una pèrdua excessiva d'informació.
- Les variables numèriques amb valors NA s'han d'imputar tal com s'indica més endavant.

A tenir en compte per fer l'activitat:

- Com a objectiu secundari, aquesta activitat pretén que desenvolupeu el coneixement del llenguatge R, aprofitant al màxim les seves característiques. En aquest sentit, us fem recomanacions de com desenvolupar el codi d'alguns apartats de manera elegant.
- Cal lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure: el codi i el resultat de l'execució del codi (pas a pas).
- Per facilitar la correcció, els fitxers s'han de lliurar per separat a l'aula. És a dir, es recomana no pujar un fitxer comprimit sinó cada fitxer per separat.
- Cal respectar la mateixa numeració dels apartats que l'enunciat.
- No es poden realitzar llistats complets del conjunt de dades a la solució. Això generaria un document amb centenars de pàgines i dificulta la revisió del text. Per comprovar les funcionalitats del codi sobre les dades, es poden fer servir les funcions **head** i **tail** que només mostren unes línies del fitxer de dades.

- Es valora la precisió dels termes utilitzats (cal utilitzar de manera precisa la terminologia de l'estadística).
- Es valora també la concisió a la resposta. No es tracta de fer explicacions gaire llargues o documents molt extensos. Cal explicar-ne el resultat i argumentar la resposta a partir dels resultats obtinguts de manera clara i concisa.

---

1 Lectura de dades i examinació del tipus de variable

1.1 Carregar el fitxer de dades

Llegir el fitxer de dades i consultar el nom de les columnes del fitxer.

1.2 Examinar el tipus de dades

Indicar quines variables són de naturalesa numèrica, caràcter i categòrica. En cas que el tipus de variable que ha atorgat R no coincideixi amb el tipus que li correspondria, indicar de quines variables es tracta. Considereu que les variables binàries prenen valors 1 o 0. La transformació corresponent, si és necessària, s'aplicarà en els apartats següents, una vegada normalitzades les variables.

---

2 Normalització de variables qualitatives (text)

2.1 Variable raceeth

Mostreu les categories de la variable raceeth. En cas d'inconsistències o errors, corregiu la informació. A continuació, mostreu el percentatge d'estudiants a cada categoria i dibuixeu un gràfic circular (pie chart).

---

3 Normalització i descripció de variables binàries

El conjunt de dades conté un nombre elevat de variables binàries. Reviseu els seus valors i en cas d'errors o inconsistències, corregiu els valors a partir dels criteris indicats. A continuació, resumiu en una taula la proporció d'estudiants per als valors positius (1) i els valors negatius (0) d'aquestes variables. Interpreteu breument.

Requisits:

- La taula ha de contenir una variable a cada fila i quatre columnes: nombre d'estudiants amb valor 0 a la variable, nombre d'estudiants amb valor 1, proporció d'estudiants amb valor 0 i proporció d'estudiants amb valor 1.
- Es recomana generar la taula de forma automàtica, sense haver de fer el càlcul manualment per a cada variable. Podeu fer servir funcions de la família *apply* per automatitzar aquest càlcul.

---

4 Normalització de variables quantitatives

4.1 Variable readingScore

Reviseu els valors de la variable readingScore i verifiqueu que estiguin dins dels marges esperats. Si hi ha algun valor erroni o molt extrem, substituir per NA. Mostreu un gràfic de tipus boxplot per visualitzar la distribució d'aquesta variable. Interpreteu el resultat.

4.2 Variable grade

Mostreu visualment la distribució de la variable grade (curs). A continuació, reviseu si els valors de la variable grade estan dins dels marges raonables. Per a la mostra d'estudi, composta per estudiants de 15 anys, es correspondria al desè curs. Hi poden haver casos d'estudiants que estiguin en cursos més avançats o en cursos inferiors. Si hi ha un valor extrem o erroni, s'ha de substituir per NA.

4.3 Variable schoolSize

Mostreu visualment la distribució de la variable schoolSize. Si hi ha valors erronis, substituïu per NA. La imputació es farà més endavant.

---

5 Imputació

En aquest apartat, farem la imputació sobre els valors perduts de la variable schoolSize. Apliqueu imputació per veïns més propers, utilitzant la distància de Gower, considerant en el còmput dels veïns més propers les variables numèriques. Per realitzar aquesta imputació, es pot fer servir la funció "kNN" de la llibreria VIM amb un nombre de veïns igual a 5. Demostreu que la imputació s'ha realitzat correctament, visualitzant algunes de les dades afectades per la imputació.

Finalment, analitzeu des d'un punt de vista crític el procés d'imputació realitzat.

---

6 Mesures característiques de les variables numèriques

Calculeu les mesures de tendència central i dispersió, tant robustes com no robustes, de les variables quantitatives numèriques grade, minutesPerWeekEnglish, studentsInEnglish, schoolSize i readingScore. Es presentaran dues taules, una amb les mesures de tendència central i una altra amb les mesures de dispersió. A la taula de tendència central, mostreu la mitjana, mediana, i mitjana retallada al 5%. A la taula de dispersió, mostreu la desviació estàndard, el rang interquartílic i la desviació absoluta respecte de la mitjana.

Requisits:

- Igual que anteriorment, realitzeu aquest càlcul sense haver de calcular la informació de cada variable per separat. Feu servir les funcions de la família *apply*.
- Per practicar el desenvolupament de funcions en R, us demanem que implementeu la funció que calcula la desviació estàndard i feu servir aquesta funció en lloc de la funció *sd* que proporciona R.

Nota:

- Com que no hem realitzat imputació per a totes les variables numèriques, si existeixen NAs en algunes variables, podeu ignorar aquests valors per fer aquests càlculs. Podeu fer servir el paràmetre *na.rm=TRUE*.

---

7 Arxiu final

Un cop realitzat el preprocessament sobre el fitxer, desar el resultat de les dades en un fitxer csv anomenat pisa_clean.csv.

Puntuació de l'activitat

- Apartat 1 (10%)
- Apartat 2 (10%)
- Apartat 3 (20%)
- Apartat 4 (20%)
- Apartat 5 (20%)
- Apartat 6 (10%)
- Qualitat de l'informe dinàmic (10%)