ביג דאטה, תרגיל 2, אביב 2024

גרסא 1.00

הגשה ביחידים

תאריך הגשה: יום שלישי, 12/7/2021 עד שעה 11:59 בלילה

בתרגיל זה נבדוק נתונים דמוגרפים בשתי מדינות: ישראל ואפגניסטן.

data

הורידו את הקובץ Life_Expectancy_Data.csv מהמודל. הקובץ מכיל מגוון נתונים דמוגרפים על מדינות בעולם לאורך 16 שנים, בין השנים 2000 ל-2016.

ישראל

צרו DataFrame שמכיל רק את הנתונים עבור ישראל. הציגו את ערכי העמודה Life.expectancy ביחס ל-Year באיור. על סמך האיור הסירו מהטבלה שורות שבהן הערכים בעמודה Life.expectancy נראים שגויים. כעת צרו מודל רגרסיה לינארית של תוחלת החיים ביחס לשנה. האם המודל מתאר יחס לינארי בין השנה לתוחלת החיים? הדפיסו למסך את ה-P-value של המודל כולו וכן את R^2.

הוסיפו קוד שמדפיס את תוחלת החיים הצפויה בשנת 2050 (בהנחה שהמגמה הנוכחית תמשך), ואת מספר החודשים שנוספים בכל שנה לתוחלת החיים. לסיום, הציגו איור שבו מוצגות הנקודות, קו הרגרסיה הלינארית וה-residuals.

אפגניסטן

- צרו DataFrame שמכיל רק את הנתונים עבור אפגניסטן. הסירו את העמודות Country ו-Status.
- עיברו על כל אחת מהעמודות ובדקו האם קיימים ערכים שגויים או חסרים בעמודה. נגדיר ערכים שגויים כערכים שלחלוטין לא מתאימים למגמה הכללית של העמודה (למשל ערך שגבוה משמעותית מהערכים שלפניו ואחריו). לצורך מעבר על העמודות אתם יכולים לצייר את הערכים ביחס לשנה, או לבחור בדרך אחרת שנראית לכם.
  - במידה וקיימים יותר משני ערכים שנראים לא תקינים, מחקו את העמודה
  - אחרת, תקנו את הערכים בצורה שנראית לכם הטובה ביותר. למשל: על ידי החלפת בממוצע, יצירת מודל לינארי על סמך שאר הערכים וכו
- צרו מודל רגרסיה לינארית מרובה נתונים שמשתמש בכל העמודות בשביל לנבא את תוחלת החיים באפגניסטן. האם המודל מצליח לנבא את תוחלת החיים בצורה טובה? אם כן אז איזה מהמשתנים הוא המשפיע ביותר? הדפיסו את המידע על כל המשתנים (שיפוע מובהקות סטטיסטית וכו).

הערות

- הגישו את כל הקוד שבו השתמשתם לצורך ביצוע המשימות. הוסיפו הערות שמסבירות מדוע ביצעתם פעולות מסויימות (למשל: תיקון ערכים) ולמה ביצעתם אותן כפי שביצעתם.
- הקפידו על איכות הקוד.

הגשה

הגישו את הקוד בקובץ בשם ex2.py דרך המודל. אין צורך לכווץ ב-zip או להגיש משהו נוסף.

בהצלחה!

- country (Nominal) - the country in which the indicators are from (i.e. United States of America or Congo)
- year (Ordinal) - the calendar year the indicators are from (ranging from 2000 to 2015)
- status (Nominal) - whether a country is considered to be 'Developing' or 'Developed' by WHO standards
- life.expectancy (Ratio) - the life expectancy of people in years for a particular country and year
- adult.mortality (Ratio) - the adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population); if the rate is 263 then that means 263 people will die out of 1000 between the ages of 15 and 60; another way to think of this is that the chance an individual will die between 15 and 60 is 26.3%
- infant.deaths (Ratio) - number of infant deaths per 1000 population; similar to above, but for infants
- alcohol (Ratio) - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
- percentage.expenditure (Ratio) - expenditure on health as a percentage of Gross Domestic Product (gdp)
- hepatitis.b (Ratio) - number of 1 year olds with Hepatitis B immunization over all 1 year olds in population
- measles (Ratio) - number of reported Measles cases per 1000 population
- bmi (Interval/Ordinal) - average Body Mass Index (BMI) of a country's total population
- under-five.deaths (Ratio) - number of people under the age of five deaths per 1000 population
- polio (Ratio) - number of 1 year olds with Polio immunization over the number of all 1 year olds in population
- total.expenditure (Ratio) - government expenditure on health as a percentage of total government expenditure
- diphtheria (Ratio) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds
- hiv/aids (Ratio) - deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births
- gdp (Ratio) - Gross Domestic Product per capita
- population (Ratio) - population of a country
- thinness.10-19.years (Ratio) - rate of thinness among people aged *10-19*
- thinness.5-9.years (Ratio) - rate of thinness among people aged 5-9
- income.composition.of.resources (Ratio) - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- schooling (Ratio) - average number of years of schooling of a population