**Assignment 2: Data Preprocessing  
HINF 5770 – Introduction to Health Data Analytics**

**Medical Condition:** Asthma

Varsha Sane EUID: vs1625

Sai Shivani Parimi EUID: sp3403

Sunitha Byrapaka EUID: sb3073

Ria Singh EUID: rs1711

**Assignment 2: Data Preprocessing**

**Q1) Open the file (asthma\_data\_2016.csv) by using pandas' read\_csv() method.**

**Answer:** The file asthma\_data\_2016.csv has been successfully loaded using the pandas read\_csv() method.

A screenshot of a computer

Description automatically generated

**Q2)** **What are the number of rows (tuples) and columns (attributes) of this file?**

**Answer:** The file contains 2,254 rows (tuples) and 1,959 columns (attributes)

A white screen with text

Description automatically generated with medium confidence

**Q3) What is the value of row index 1494 and column index 229?**

**Answer:** The value of row index 1494 and column index 229 is 4

A screen shot of a computer

Description automatically generated

**Q4)** **Select the following variables: ["DUPERSID", "SEX", "ASPRIN53", "ADAPPT42", "ADHECR42", " AGE16X", "BMINDX53", "CHBMIX42", "FAMINC16", "WAGEP16X", "TTLP16X", "UNEMP16X", "RACETHX", "TOTEXP16"]. Then, assign the selected attributes to a new data called asthma\_selected.**

**Answer:** This will create the new asthma\_selected dataframe with the chosen columns.

A screenshot of a computer

Description automatically generated

**Q5) Show (i.e., print) the values of DUPERSID and SEX in rows from 150 to 160. Here, numbers indicate row indexes.**

A screenshot of a computer

Description automatically generated

**Q6) By using describe method (), examine the distribution of data and answer two questions.**

A screenshot of a computer

Description automatically generated

**6a: Which attribute does seem to have the most missing values? (5 points)**

**Answer:** CHBMIX42 has the most missing values.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

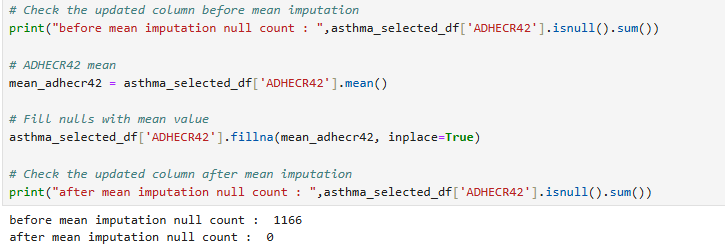
**6b: List the names of categorical variables. Explain why you think they are categorical variables. (5 points)**

**Answer**: SEX: As it represents gender and has two categories MALE and FEMALE.

RACETHX: Represents race/ethnicity categories.

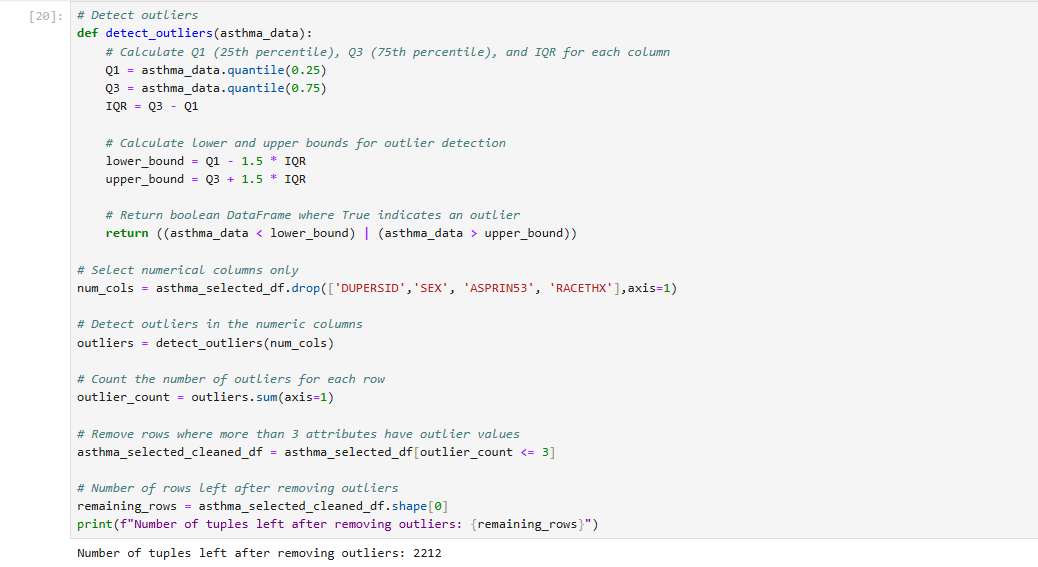
ASPRIN53: As per the code book there are two values i.e., 1, 2 which represent Yes and No respectively. So, this is considered categorical.

**Q7) Change missing values of ADHECR42 to the mean value of this attribute.**

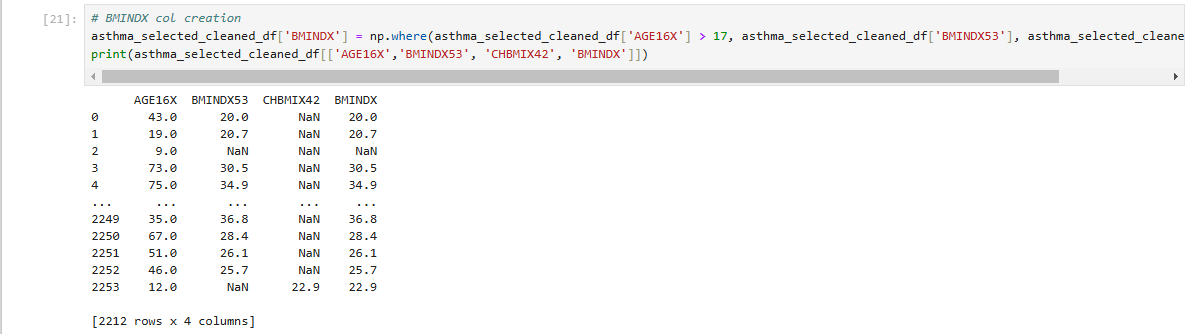


**Q8) Remove tuples if more than three attributes' values are below Q1 – 1.5×IQR or above Q3 + 1.5×IQR. In other words, remove outliers. How many tuples are left?**

**Answer:** After removing outliers total number of tuples left = 2212



**Q9) Combine BMINDX53 (for adults, >17) and CHBMIX42 (for children, 0-17) into one variable called BMINDX.**



**Q10) We need to avoid redundancy in attributes. We can use the correlation coefficient for numeric data and the Chi-square test for categorical data. After applying the correlation coefficient, what attribute would you drop? Explain your rationale.**

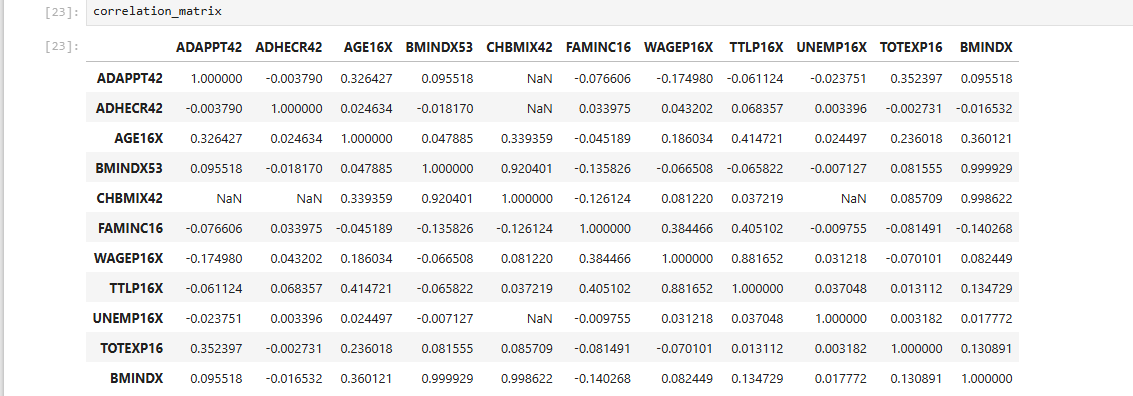
**Answer:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



* Numerical variables: Since there is high correlation between BMINDX53, BMINDX, and CHBMIX42 variables, dropping BMINDX53 and CHBMIX42 and keeping only BMIDX.
* WAGEP16X and TTLP16X also have high correlation, So, dropping WAGEP16X.

A screenshot of a computer

Description automatically generated

* Categorical Variable: The p-value between RACETHX and ASPRIN53 is 0.023 which indicates significant correlation. Considering, RACETHX gives more information on understanding asthma prevalence and severity retaining that and dropping ASPRIN53.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Q11) Apply Max-Min normalization to numerical attributes.**

A screenshot of a computer

Description automatically generated

**Q12) Save the file as “asthma\_data\_processed.csv**

A screenshot of a computer

Description automatically generated