* Create variable ‘gender’ based on Mr. or Mrs. on ‘name’
* Create variable ‘age’ based on ‘birth year’
* Create variable ‘BMI’ based on ‘weight’ and ‘height’
* ‘smoking habit’ should be binary
* ‘exercise’ should be binary
* ‘education’ has the following categories:
  + 'High School Incomplete (10th to 11th grade)'
  + ‘Elementary School (1st to 9th grade)',
  + 'University Complete (3 or more years)',
  + 'University Incomplete (1 to 2 years)',
  + 'High School Graduate', 'I never attended school / Other',
  + MISSING VALUES
* ‘drinking habit’ has the following categories:
  + 'I usually consume alcohol every day',
  + 'I consider myself a social drinker',
  + 'I do not consume any type of alcohol'
* ‘fruit habit’ has the following categories:
  + 'Less than 1. I do not consume fruits every day.',
  + '1 to 2 pieces of fruit in average',
  + '3 to 4 pieces of fruit in average',
  + '5 to 6 pieces of fruit in average',
  + 'More than six pieces of fruit'
* ‘water habit’ has the following categories:
  + 'Between one liter and two liters',
  + 'More than half a liter but less than one liter',
  + 'Less than half a liter'
* ‘checkup’ has the following categories:
  + 'More than 3 years',
  + 'Not sure',
  + 'Less than 3 years but more than 1 year',
  + 'Less than three months'
* ‘diabetes’ has the following categories:
  + 'Neither I nor my immediate family have diabetes.',
  + 'I have/had pregnancy diabetes or borderline diabetes',
  + 'I do have diabetes',
  + "I don't have diabetes, but I have direct family members who have diabetes."
* ‘region’ has the following categories:
  + 'London',
  + 'South West',
  + 'Yorkshire and the Humber',
  + 'South East',
  + 'East Midlands',
  + 'West Midlands',
  + 'East of England',
  + 'North West',
  + 'North East',
  + 'LONDON' 🡪 problem, because should be the same as London.

**FAZER:**

* One-hot encoder to the categorical variables.
* Solve problem with outliers ‘age’ and ‘blood pressure’
* Solve problem with LONDON
* Feature selection
* Model selection:
  + Logistic regression
  + Naive bias