Heart Disease Prediction

Vikas Ranjan

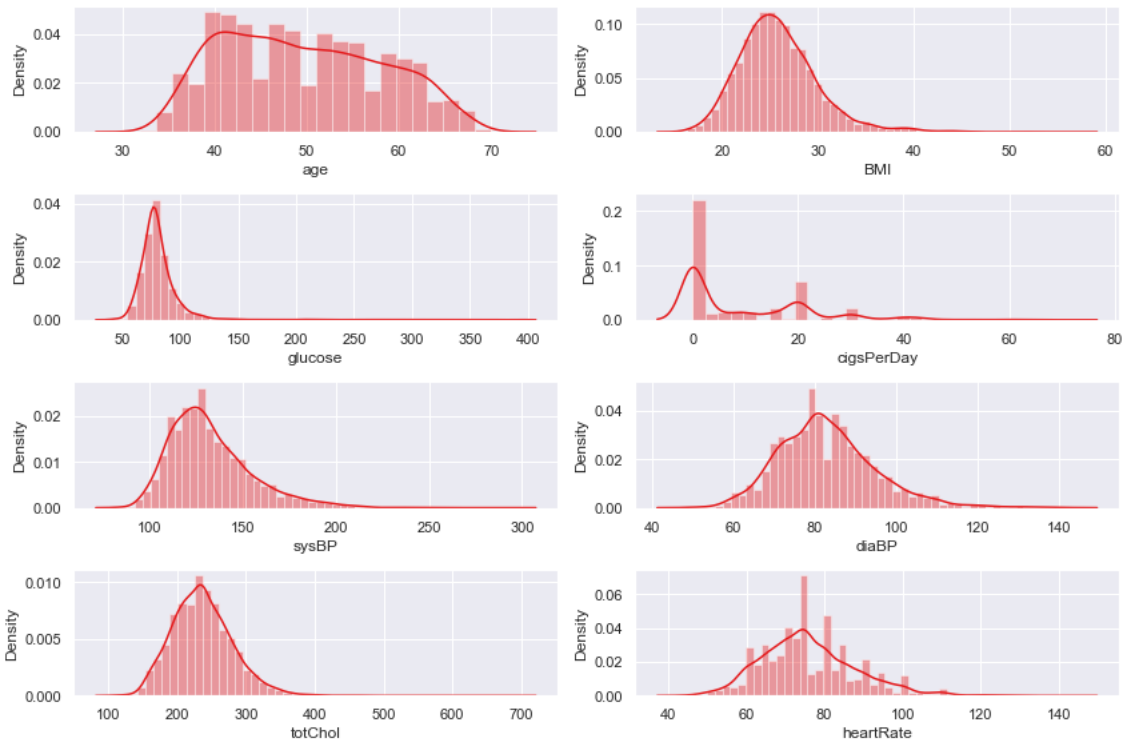
DSC680, Summer 2021

Bellevue University, NE

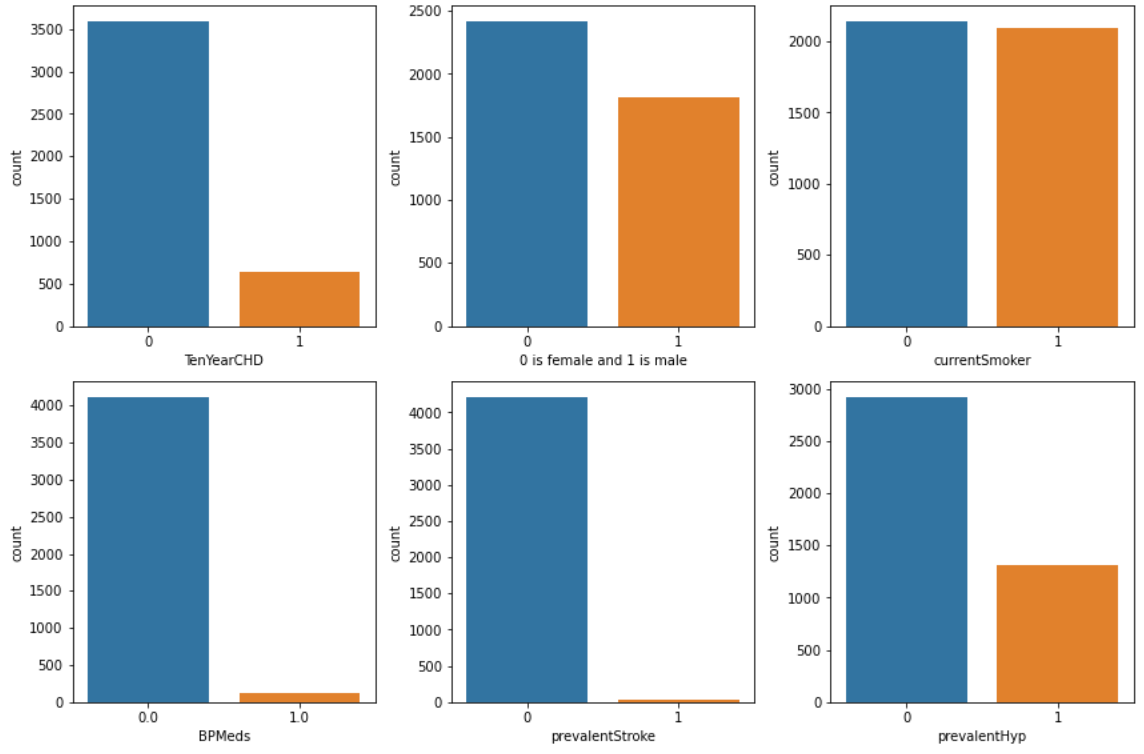
Objective/Questions:

1. Identify and analyze distribution of key risk factors with regards to risk of coronary heart disease?

   - Below is the distribution of all the continuous attributes in the dataset.

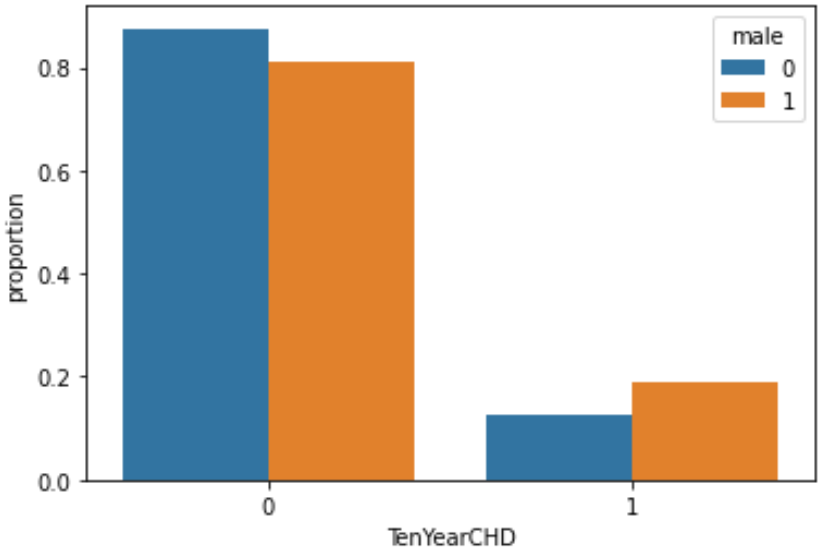


   - Below is the distribution of all the boolean attributes in the dataset.



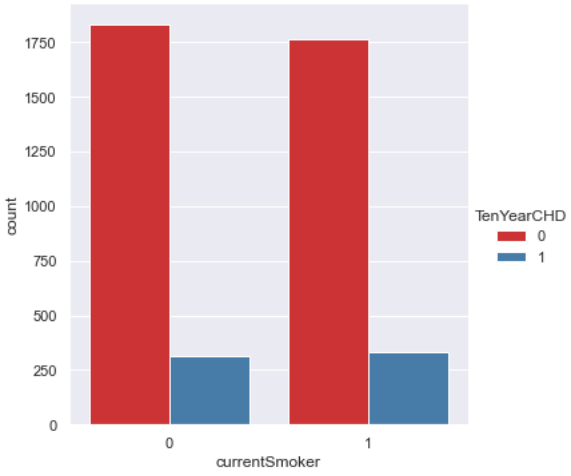
2. Which gender has more risk of coronary heart disease (CHD)?

   - As we can see below, most of the people who don't have a 10-year risk of coronary heart disease are female while the ones who have the risk are generally male.



3. Identify and plot the impact of smoking status to risk of heart disease?

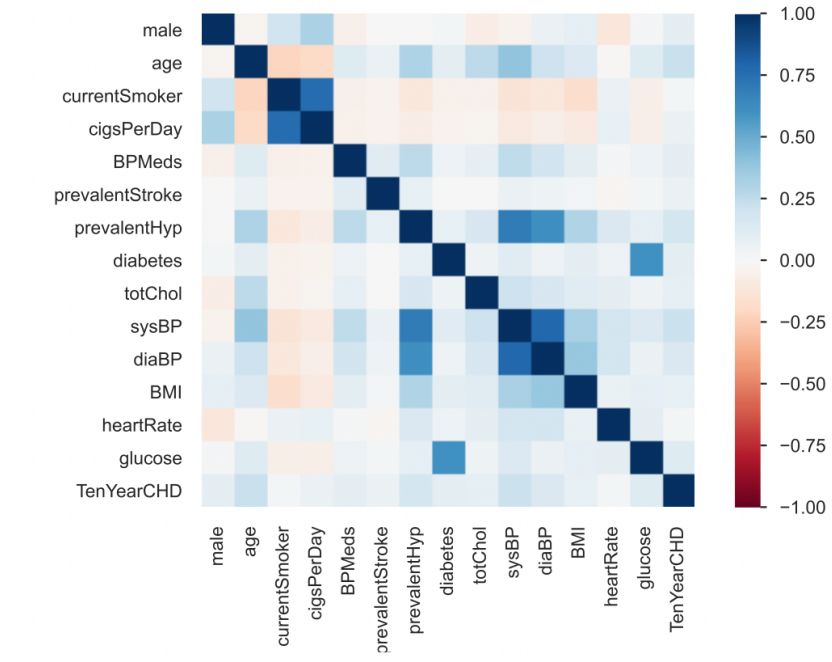
- We can see in below plot; smoking status has insignificant effect on the risk in the data.

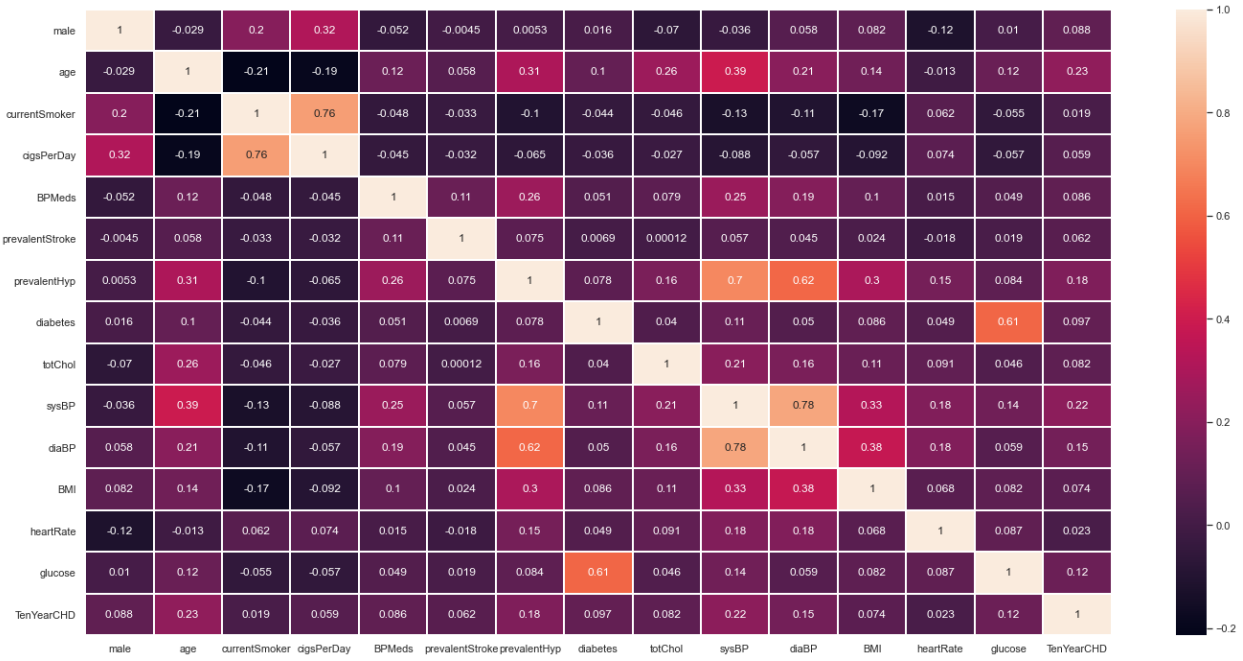


4. Find out which risk factors are highly corelated with regards to risk of heart disease?

- Heatmap of the correlation.

Pearson Correlation:





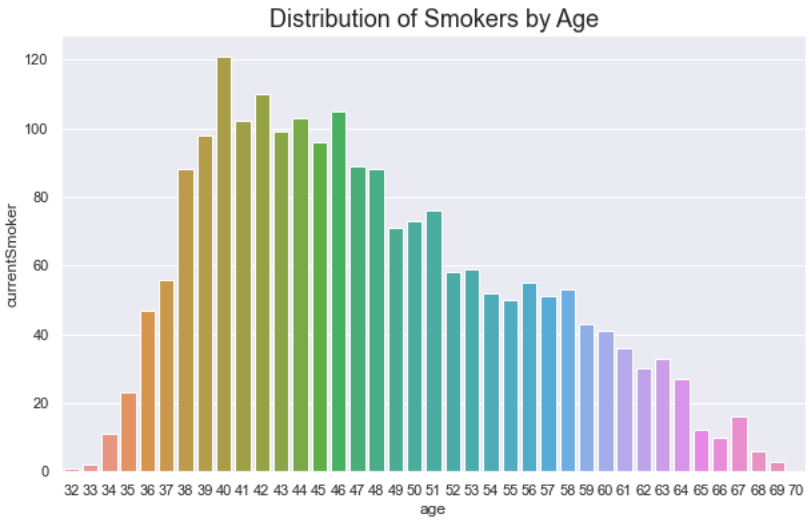

```
TenYearCHD          1.000000
age                 0.225256
sysBP               0.216429
prevalentHyp        0.177603
diaBP               0.145299
glucose             0.121277
diabetes            0.097317
male                0.088428
BPMeds              0.086417
totChol             0.081566
BMI                 0.074217
prevalentStroke     0.061810
cigsPerDay          0.058859
heartRate           0.022857
currentSmoker       0.019456
Name: TenYearCHD, dtype: float64
```

Correlations Analyis - TenYearCHD:

- The strongest positive correlations of TenYearCHD are with age and sysBP.
- CurrentSmoker has the least correlation with TenYearCHD.

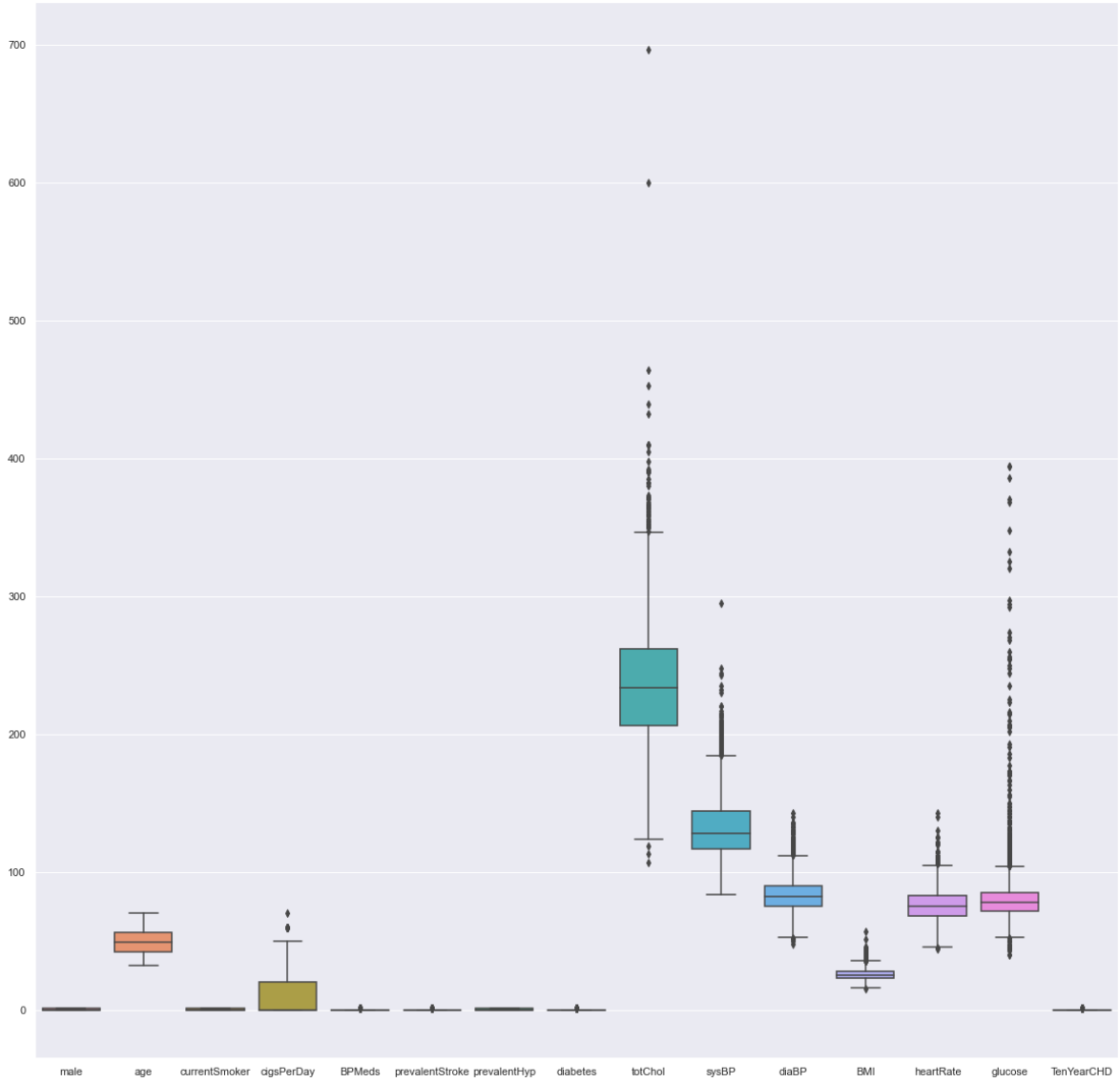
5. Identify the distribution of smokers by age?

- Below plot shows that highest number of smokers are in age range of 40-46.

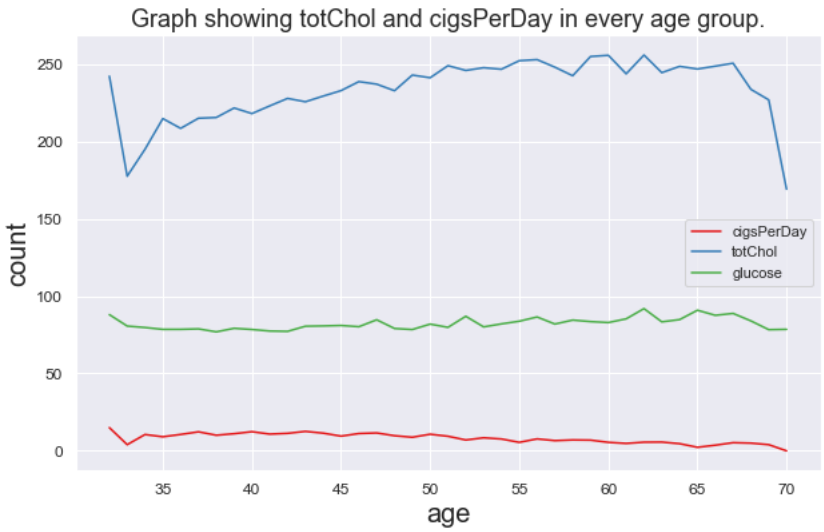


6. Identify if there are any outliers in the dataset and determine how to handle them?

   - Below box and whisker plot shows that totChol and sysBP has outliers. I'm going to remove the records which have totChol more than 600 and sysBP more than 295.

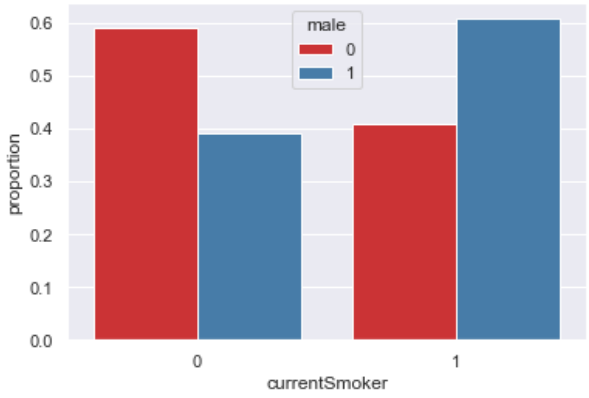


7. Plot and analyze risk factors cigsPerDay, totChol, glucose levels by age.



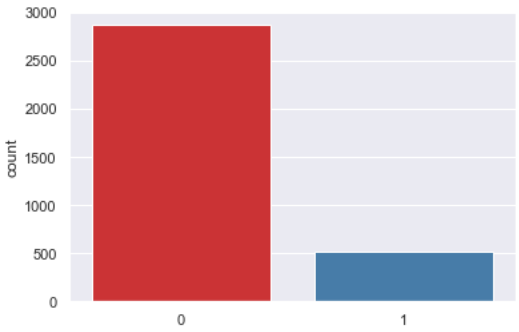
8. Identify the percentage of people smokers by gender.

   - Below graph shows the proportion of people in the status smokers and non-smokers while the colors show the gender in each group. In this data, almost 60 percent of people who are smokers are female while more than 60 percent of people who are not smokers are male.



9. Validate if the dataset is imbalanced for applying a classification model? If yes, select a method to overcome this challenge.

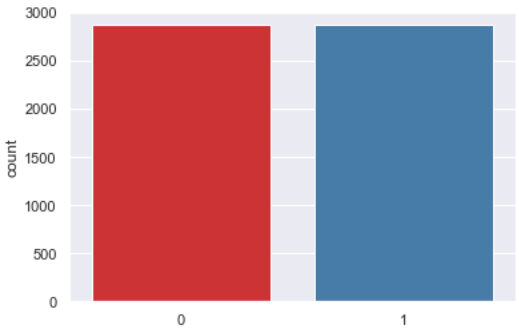
   - Comparison of records with TenYearCHD values as 0 (No) and 1 (Yes).



   -

   As we can see above, this is a moderately imbalanced dataset. Therefore, we would need to handle that using random oversampling.

   - Distribution of the training dataset after oversampling.



10. Identify which classification model will be best suited to predict the risk of coronary heart disease?

    - Below is the comparison of the implemented models in terms of accuracy. Here we can see that, based on the accuracy of the implemented classification models to predict TenYearCHD, SVM classification model shows highest accuracy, followed by Random Forest classification.

	Model	Score
1	Random Forest Classifier	0.937789
3	Decision Tree Classifier	0.929115
2	KNN	0.816730
0	Logistic Regression	0.680041