AT THE HEART OF THE MATTER

AN EXTENSION OF THE CLEVELAND CLINIC'S ANALYSIS OF HEART DISEASE INDICATORS

BACKGROUND

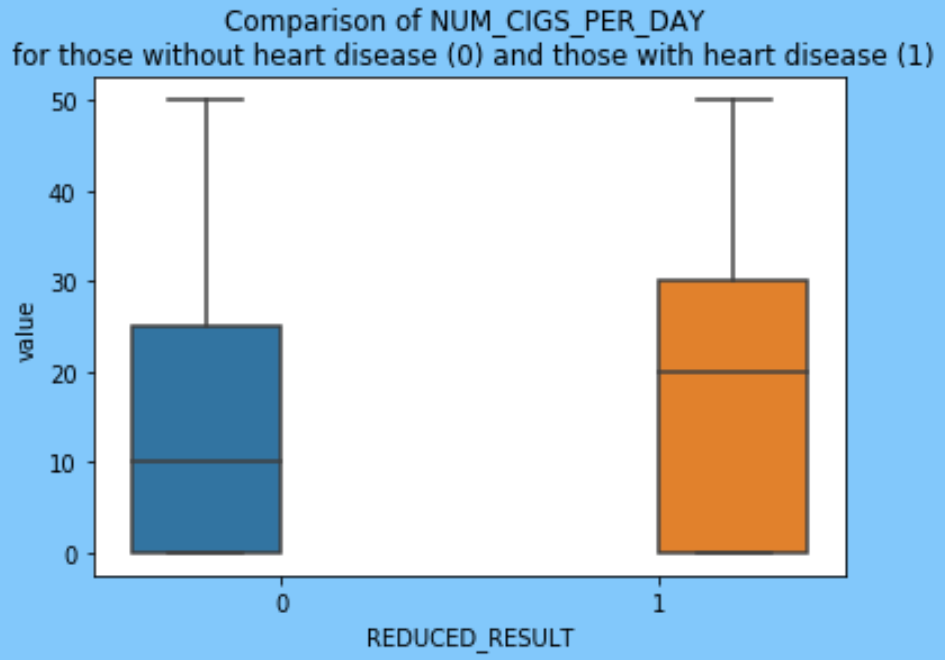
- Heart disease deaths increased by 3% between 2011 and 2014.
- The patient's age, gender, fasting blood sugar, and blood cholesterol have been identified as indicators of heart disease.
- In this study, I examine how a patient's heart rate, blood pressure, and smoking history act as indicators of heart disease. I chose these indicators because they have not been previously examined and they are easily measured.

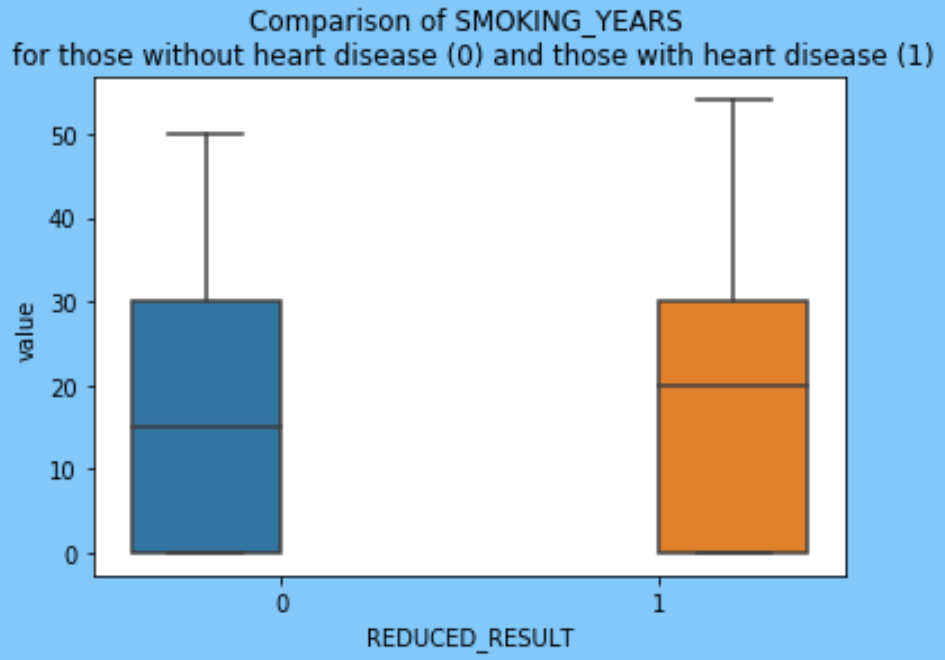
EXPLORATORY DATA ANALYSIS

- An initial examination of the dataset revealed nearly 900 records with 76 fields per record.
- It also showed that a significant amount of data consisted of inappropriate values. Some examples of these included:
    - The *heart rate* and/or *blood pressure* fields were null or empty.
    - The *is smoker* field, the *number of cigarettes per day*, and the *years smoked* were either incongruent or were null.
    - The *history of heart disease* field was not either a zero or a one.
- Once the data was cleansed, roughly 400 records remained with 8 fields per record.

EXPLORATORY DATA ANALYSIS

- Below are comparisons showing the distributions of the number of cigarettes smoked per day and the number of years of smoking for those with heart disease and those without heart disease.





EXPLORATORY DATA ANALYSIS

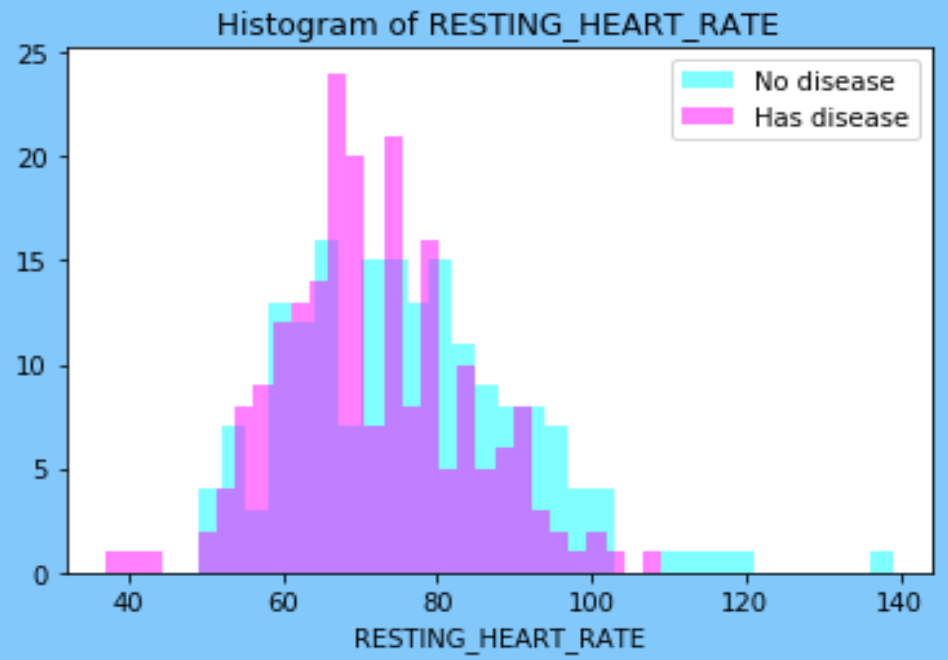
- Comparison of statistics for the previous plots.

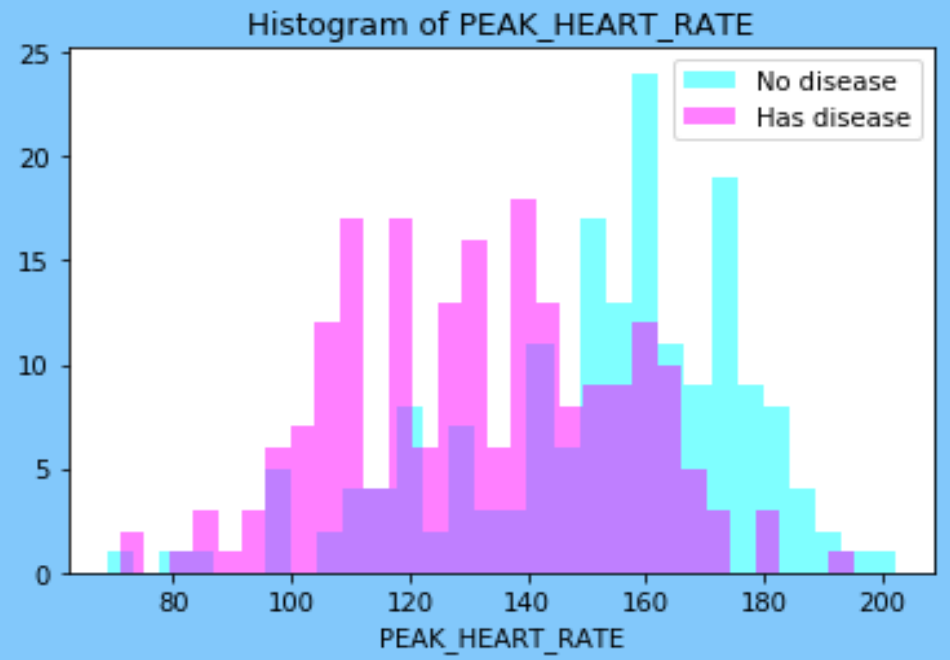
Sample Set	Difference in Means	Margin of Error	P-value
Num. of Cigs / Day	-25.43	1.46	1.000
Smoking Years	-4.12	3.23	0.993

- These were calculated via permutation tests assuming that the difference in means of the two sample populations (smokers with heart disease and smokers without heart disease) was equal to zero. Given that the p-values are so high, we are unable to reject the null hypothesis. In other words, there is no statistical evidence that, in this population, smoking is an indicator of heart disease.

EXPLORATORY DATA ANALYSIS

- Comparisons showing resting and peak heart rates for those with heart disease and those without heart disease.





EXPLORATORY DATA ANALYSIS

- Comparison of statistics for the previous plots.

Sample Set	Difference in Means	Margin of Error	P-value
Resting Heart Rate	4.25	2.74	< 0.05
Peak Heart Rate	20.75	4.81	< 0.05

- These were calculated using permutations tests assuming that the difference in means of the two sample populations (those with heart disease was and those without heart disease) was equal to zero. Given that the p-values are so low, we may reject the null hypothesis. In other words, there is statistical evidence that, in this population, there is a measurable difference in resting and peak heart rate for those with heart disease and those without heart disease.

MACHINE LEARNING

- We chose a logistic regression model because it allows us to more easily interpret the results.
- After mitigating issues with the indicators' multicollinearity, I found that the following indicators and coefficients provided a best fit[1] for the data, with an accuracy of 0.76:

Indicator	Coefficient
Number of cigarettes smoked / day	0.007
Years of Smoking	0.006
Peak Heart Rate	-0.033
Peak Diastolic Blood Pressure	0.009

Best fit can be understood as a compromise between the removing indicators which complicate the model and keeping enough indicators to make the model useful.

MACHINE LEARNING

- This leads to the following equation, which can be used to calculate the probability that someone has heart disease, where
  - x_1 = the number of cigarettes smoked per day
  - x_2 = the number of years of smoking
  - x_3 = the peak heart rate
  - x_4 = the peak diastolic blood pressure
- $p = \frac{1}{1+e^{-(0.007x_1 + 0.006x_2 - 0.033x_3 + 0.009x_4)}}$

MACHINE LEARNING

- Here is an example of how to use this expression. Consider a patient how smokes 10 cigarettes per day and has done so for 20 years. The patient has a peak heart rate of 120 beat per minute and a peak diastolic pressure of 80 mm Hg. We can use the expression to calculate the probability that the person has heart disease as follows:
- $p = \frac{1}{1+e^{-(0.007*10 + 0.006*20 - 0.033*120 - 0.009*80}} = 0.05$
- Here we see that there is a 5% probability that the person will have heart disease.

CONCLUSION

- Selected a dataset and explored it.
- Used logistic regression to model the relationship between the probability of having heart disease and several potential indicators.
- Based on the model, I recommend that if a medical practitioner notices a statistically significant drop in a person's peak heart rate, they test specifically for heart disease.
- This model is limited and there is much opportunity to expand on this effort. A larger and more complete data set could provide a more accurate model. Also, studying how indicators may influence one another may prove beneficial when health measurements are made.