Idan Chen

Machine Learning project – Phase 1

Dataset - Personal Key Indicators of Heart Disease

1) **Q**: What are the business needs you are trying to address

   A: My business needs I'm trying to address is to understand how adult people can improve their health status

2) **Q**: What are the objectives of the Data Science project – use supervised learning (regression or classification) and define how your target, y, addresses the business need/problem

   A: my y target is "HeartDisease" and I will predict the result (0/1) based on my features and when I will know the weight of each feature I will understand how adult people can Improve the chance of not getting heart disease.

3) **Q**: What kind of data is available

   A: The dataset is about " Key Indicators of Heart Disease" based on " 2020 annual CDC survey data of 400k adults related to their health status" and has columns about heart disease (the target), BMI, Smoking, Alcohol Drinking, Age Category and etc.…

4) **Q**: How large is your data

   A: My data has 400K rows and 18 columns

5) **Q**: Main features

   A: My main features are: Smoking, Stroke, AgeCategory, PhysicalHealth, DiffWalking, Diabetic, SkinCancer and KidneyDisease

6) **Q**: Can you find a new data set online that you could merge and increase your insights

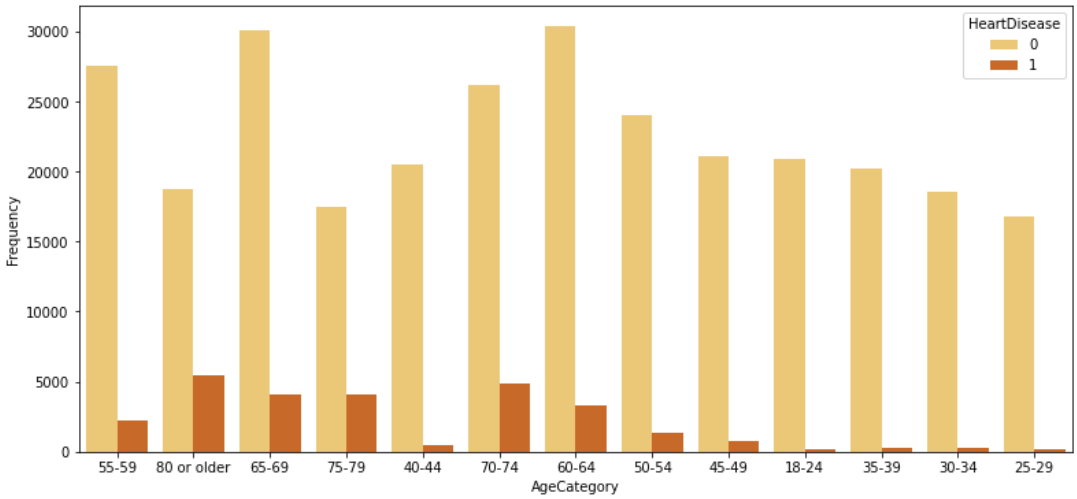
A: Yes, I took my data from Kaggle, the name of the dataset is "Personal Key Indicators of Heart Disease"

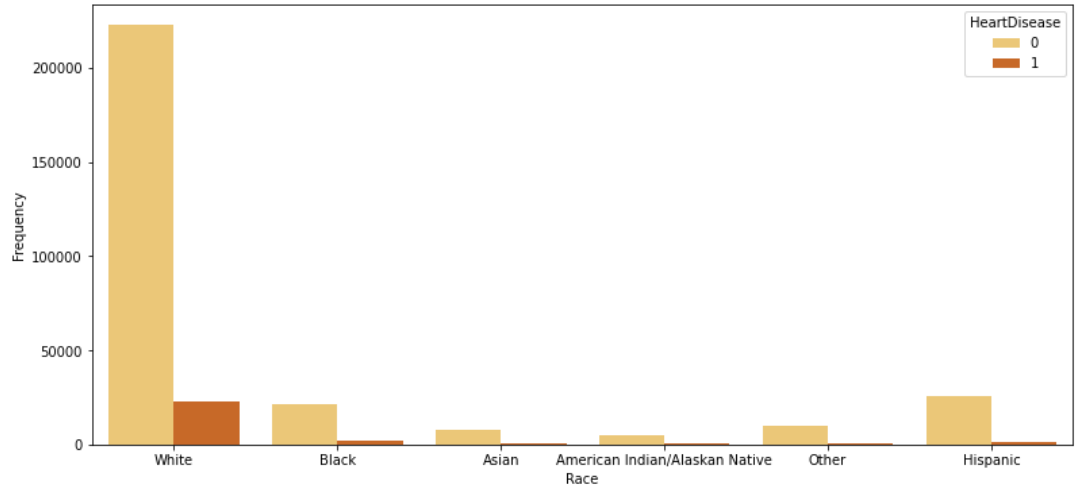
The original location came from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents.

I can go to their articles and get more data from their surveys.

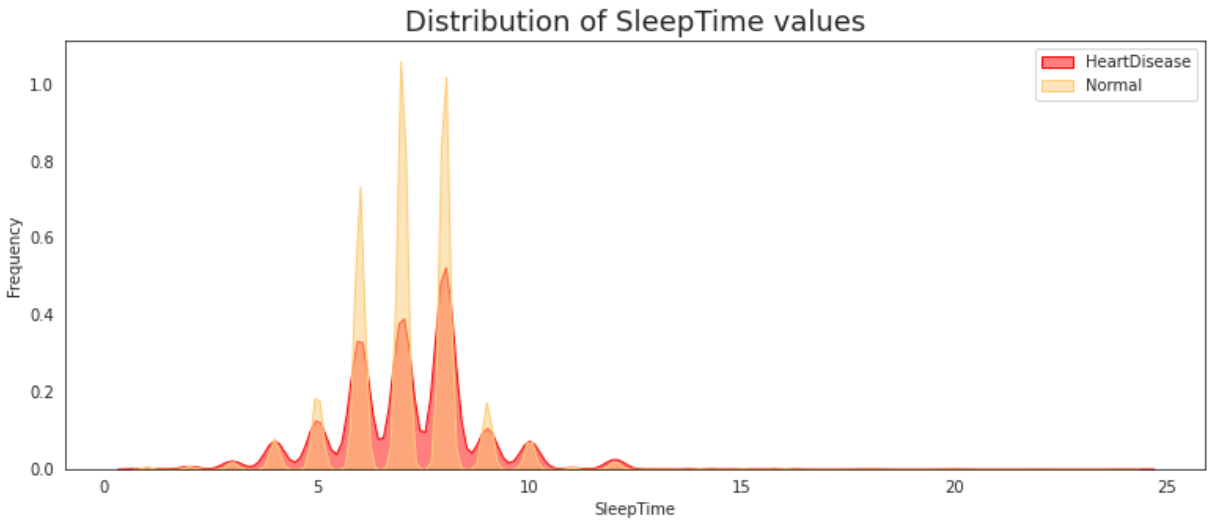
7) **Q**: Exploratory Data Analysis – 2-6 interesting visualizations. Self-explanatory

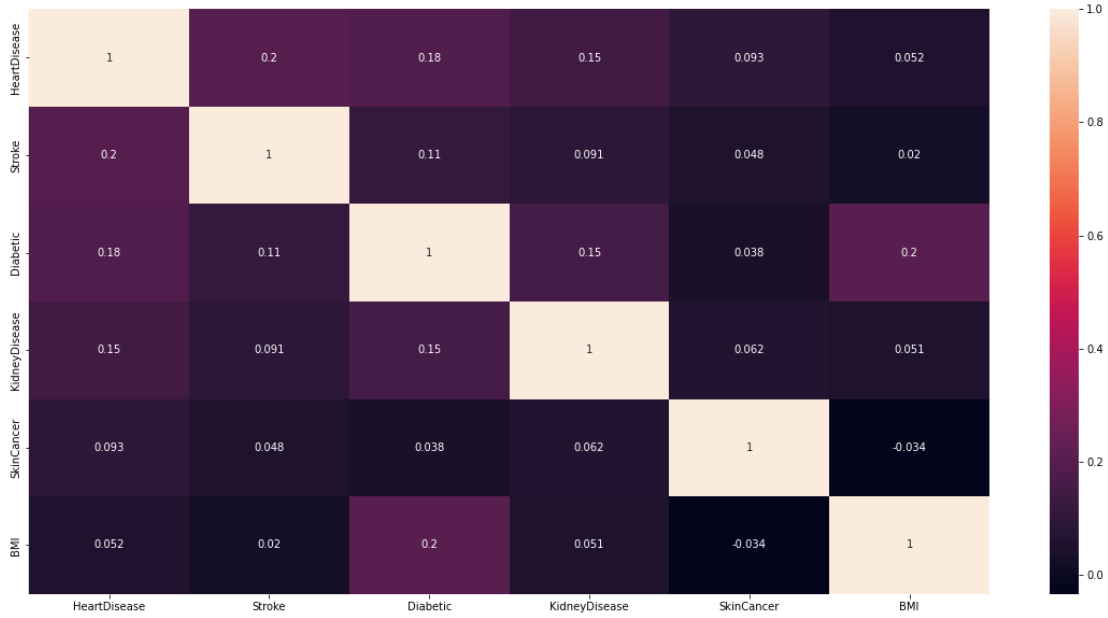
A:



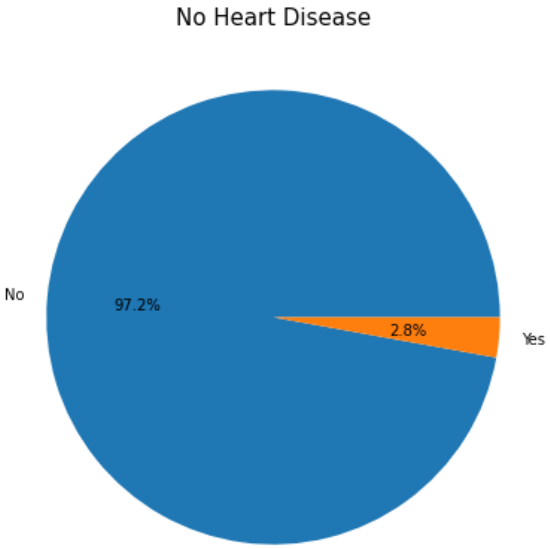


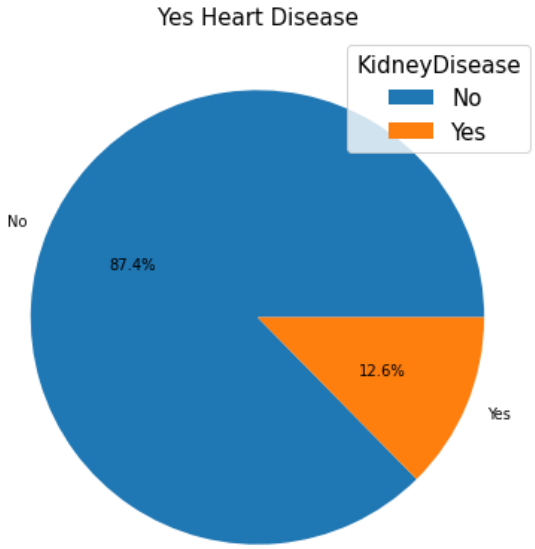
Idan Chen

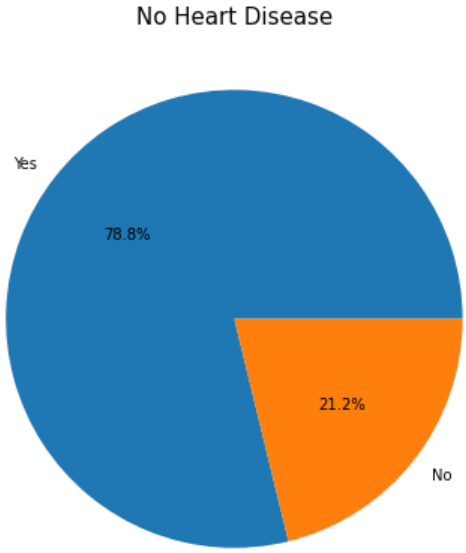

Distribution of SleepTime values
HeartDisease
Normal
Frequency
SleepTime
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25

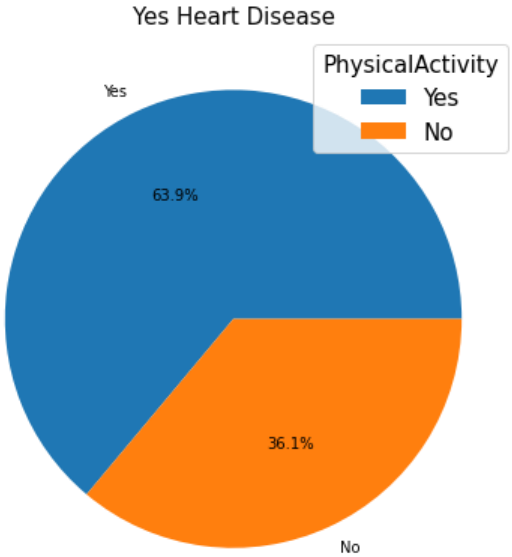

HeartDisease
Stroke
Diabetic
KidneyDisease
SkinCancer
BMI
1
0.2
0.18
0.15
0.093
0.052
0.2
1
0.11
0.091
0.048
0.02
0.18
0.11
1
0.15
0.038
0.2
0.15
0.091
0.15
1
0.062
0.051
0.093
0.048
0.038
0.062
1
-0.034
0.052
0.02
0.2
0.051
-0.034
1
1.0
0.8
0.6
0.4
0.2
0.0

Idan Chen


No Heart Disease
No
97.2%
2.8%
Yes


Yes Heart Disease
KidneyDisease
No
Yes
No
87.4%
12.6%
Yes


No Heart Disease
Yes
78.8%
21.2%
No


Yes Heart Disease
PhysicalActivity
Yes
No
Yes
63.9%
36.1%
No