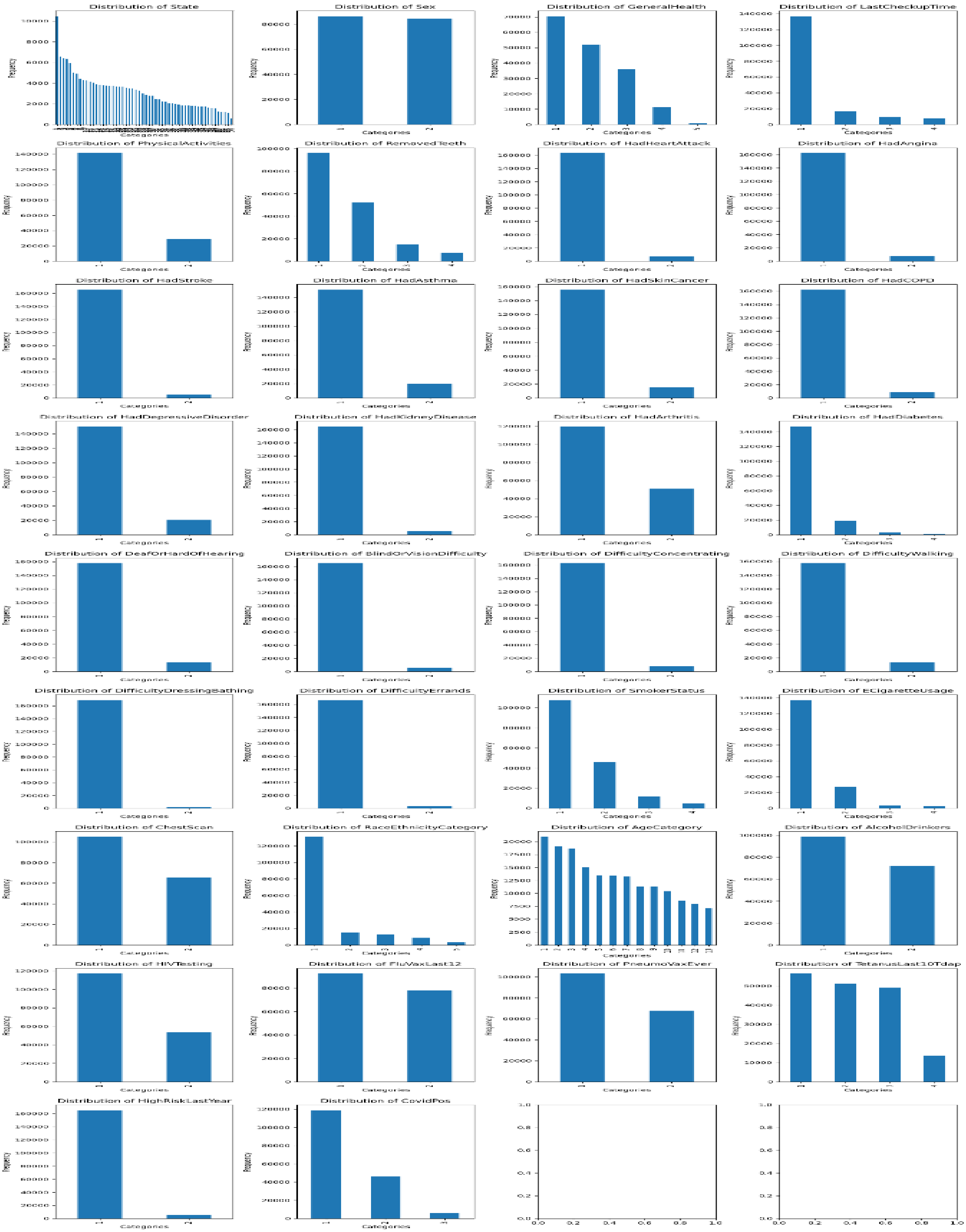
Project Part 2

Team: Group 15
Member: Yunlong Ou, Yunzhuo Liu, Yuyun Zhen, Te-Hsin Kung, Joshua Liu

1. The goal of the project is to predict when a patient is susceptible to a heart disease, and thus a heart attack. This involves investigating different variables and understanding their impact. The chosen dataset was collected by the CDC. The response variable is HadHeartAttack & HadAngina , where “Yes” indicates the patient had heart disease and “No” indicates the patient did not have heart disease. The features we plan to use are: State, Sex, GeneralHealth, PhysicalHealthlyDays, MentalHealthlyDays, LastCheckupTime, SleepHours, HadAngina, HadStroke, HadDiabetes, SmokerStatus, ECigaretteUsage, RaceEthnicityCategory, AgeCategory, HeightInMeters, WeightInKilograms, HighRiskLastYear. These features will help us determine heart disease risk by exploring the complex relationships between lifestyle factors.

2. A lot of the features are categorical variables indicating if the patient has a specific condition. We can group these into a single. For example, make a new feature called ExistingHealthIssues that counts the number of health issues such as HadStroke, HadDiabetes, etc., that a patient has. Also, some features have additional descriptions in their values. For example, ECigaretteUsage has values: [“Never used e-cigarettes in my entire life”, “'Not at all (right now)”, “Use them every day”, “'Use them some days”]. We can combine categories that we don’t need to distinguish between. For this case, we can make this a binary variable where “'Not at all (right now)” is False and the rest of the values are True. We may also create new features such as BMI (Body Mass Index) and to capture the relationship between variables like SmokerStatus and AgeCategory to capture the combined effect for heart disease. We’ll also decide whether to retain data of varying degrees under the same category based on experiments, considering factors such as relevance to heart disease and distribution of the data.

3. The Kaggle description of the dataset mentions that the classes are imbalanced. Also, our initial step of data cleaning also indicates the same result. To address this, we plan to oversample the class where HadHeartAttack = “No”. This will be useful in classification models such as decision trees, SVM, SMOTE, etc. We plan to implement SMOTE after splitting the data into training and testing sets, ensuring that we do not introduce bias into our evaluation metrics. SMOTE will help balance the classes by generating synthetic samples and ensuring our model is not biased.



Distribution of State
Frequency
Categories
0
2000
4000
6000
8000
10000
Distribution of Sex
Frequency
Categories
0
20000
40000
60000
80000
1
2
Distribution of GeneralHealth
Frequency
Categories
0
10000
20000
30000
40000
50000
60000
70000
1
2
3
4
5
Distribution of LastCheckupTime
Categories
0
20000
40000
60000
80000
100000
120000
140000
1
2
3
4
Distribution of PhysicalActivities
Frequency
Categories
1
2
Distribution of RemovedTeeth
Frequency
Categories
1
2
3
4
Distribution of HadHeartAttack
Frequency
Categories
1
2
Distribution of HadAngina
Frequency
Categories
1
2
Distribution of HadStroke
Frequency
Categories
1
2
Distribution of HadAsthma
Frequency
Categories
1
2
Distribution of HadSkinCancer
Frequency
Categories
1
2
Distribution of HadCOPD
Frequency
Categories
1
2
Distribution of HadDepressiveDisorder
Frequency
Categories
1
2
Distribution of HadKidneyDisease
Frequency
Categories
1
2
Distribution of HadArthritis
Frequency
Categories
1
2
Distribution of HadDiabetes
Frequency
Categories
1
2
3
4
Distribution of DeafOrHardOfHearing
Frequency
Categories
1
2
Distribution of BlindOrVisionDifficulty
Frequency
Categories
1
2
Distribution of DifficultyConcentrating
Frequency
Categories
1
2
Distribution of DifficultyWalking
Frequency
Categories
1
2
Distribution of DifficultyDressingBathing
Frequency
Categories
1
2
Distribution of DifficultyErrands
Frequency
Categories
1
2
Distribution of SmokerStatus
Frequency
Categories
1
2
3
4
Distribution of ECigaretteUsage
Frequency
Categories
1
2
3
4
Distribution of ChestScan
Frequency
Categories
1
2
Distribution of RaceEthnicityCategory
Frequency
Categories
1
2
3
4
5
Distribution of AgeCategory
Frequency
Categories
0
2500
5000
7500
10000
12500
15000
17500
20000
1
2
3
4
5
6
7
8
9
10
11
12
13
Distribution of AlcoholDrinkers
Frequency
Categories
1
2
Distribution of HIVTesting
Frequency
Categories
1
2
Distribution of FluVaxLast12
Frequency
Categories
1
2
Distribution of PneumoVaxEver
Frequency
Categories
1
2
Distribution of TetanusLast10Tdap
Frequency
Categories
0
10000
20000
30000
40000
50000
1
2
3
4
Distribution of HighRiskLastYear
Frequency
Categories
1
2
Distribution of CovidPos
Frequency
Categories
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0