Final Report

Will Gullion, Brian Nalley, Nate Nellis, McKade Thomas

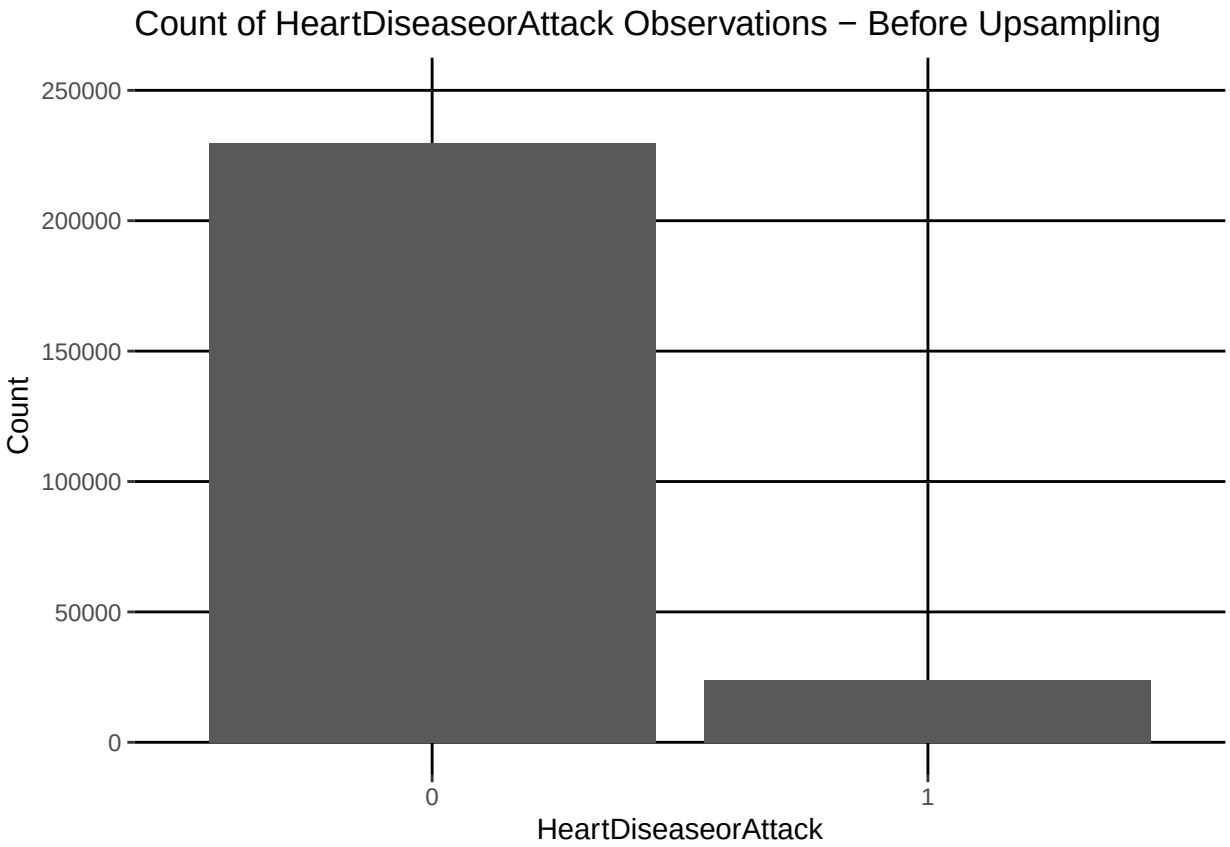
STAT 5650

Introduction

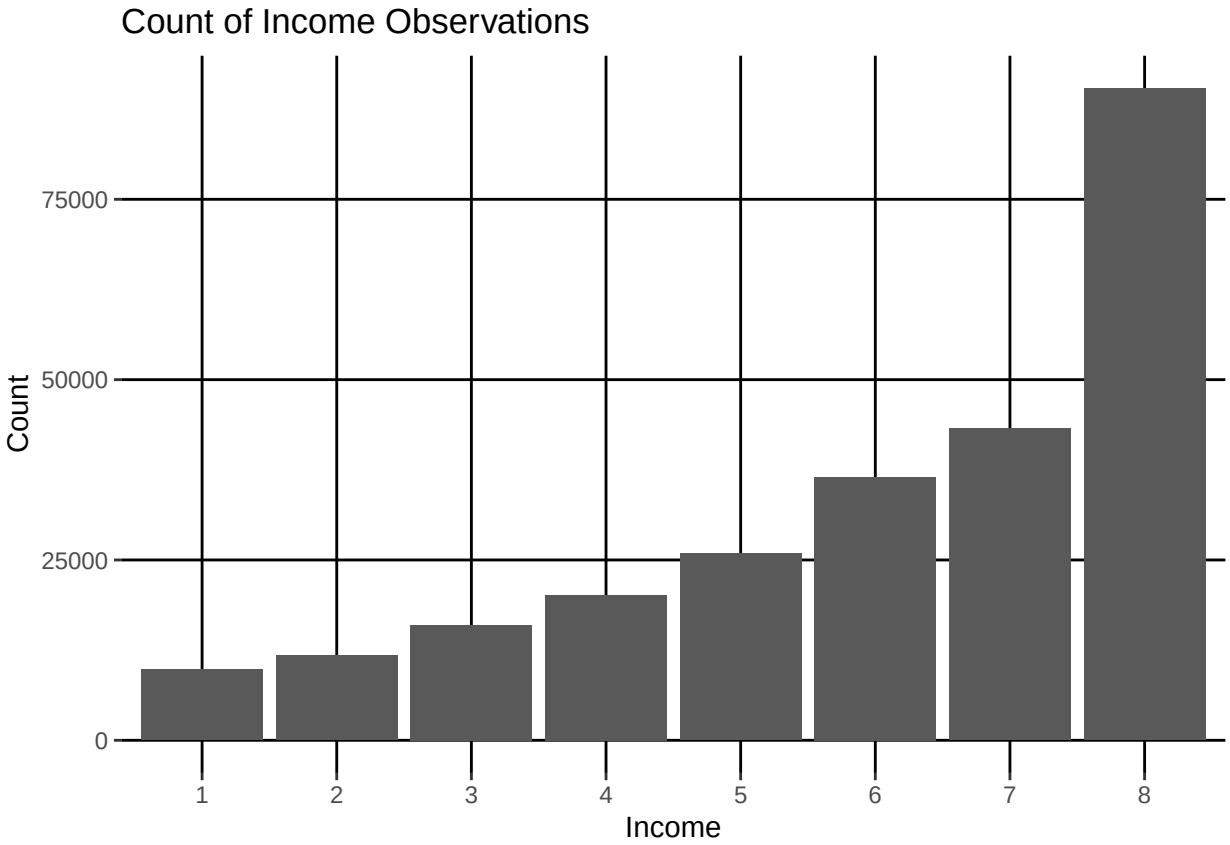
We chose to focus on data having to do with heart disease/heart attach indicators to understand which factors impact a person's probability to have a heart attach or summer from strong levels of heart disease. Data was found on Kaggle and was modified from a CDC survey. See the *Data Dictionary* section for more details at the end of the document. The dataset contains 252,680 observations. The response variable we have is a binary one telling us if the person being interviewed suffered from a heart attack or not. There are a total of 20 available predictor variables.

Exploratory Data Analysis

Taking a closer look at our predictor variable, we noticed that only 9.4% of participants in the survey had heart disease or a heart attack, 23,893 responded yes versus 229,787 who responded no. This means that we could just predict a 0 for all observations and we could get an accuracy of over 90%. Since we aren't concerned in this instance with a high accuracy, but actually a decent sensitivity (getting accurate results when someone is at risk of a heart attack or heart disease), we utilized Upsampling in some of our methods to skew the model to predict the results we wanted. This decreased overall accuracy, but resulted in a much better prediction accuracy of finding those people with heart disease or who had a heart attack.



When it comes to the predictor variables there are 12 binary, 3 numerical, and 5 factor variables. Having fewer numerical variables is less than desireable, especially when it comes to how the factor variables are coded/binned. An example of this terrible survey coding is the **Income** variable where the CDC survey sets the highest bucket capped at $75,000+. Due to how low that is for the maximum response allowed, a full third of the observations fall into that category.



Data Dictionary

Data found by us at https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

Description of variables found in https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Response Variable

HeartDiseaseorAttack, binary

```
summary(health$HeartDiseaseorAttack)
```

```
##      0      1
## 229787  23893
```

Predictor Variables

HighBP, binary
- Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

```
summary(health$HighBP)
```

```
##      0      1
## 144851 108829
```

HighChol, binary
- Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?

```
summary(health$HighChol)
```

```
##      0      1
## 146089 107591
```

CholCheck, binary
- Cholesterol check within past five years

```
summary(health$CholCheck)
```

```
##      0      1
##   9470 244210
```

BMI, numerical

```
summary(health$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   24.00   27.00   28.38   31.00   98.00
```

Smoker, binary
- Have you smoked at least 100 cigarettes in your entire life?

```
summary(health$Smoker)
```

```
##      0      1
## 141257 112423
```

Stroke, binary

```
summary(health$Stroke)
```

```
##      0      1
## 243388  10292
```

Diabetes, factor
- 0 is no diabetes,
- 1 is pre-diabetes
- 2 is diabetes

```
summary(health$Diabetes)
```

```
##      0      1      2
## 213703   4631  35346
```

PhysActivity, binary
- During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?

```
summary(health$PhysActivity)
```

```
##      0      1
##  61760 191920
```

Fruits, binary
- Consume Fruit 1 or more times per day

```
summary(health$Fruits)
```

```
##      0      1
##  92782 160898
```

Veggies, binary
- Consume Vegetables 1 or more times per day

```
summary(health$Veggies)
```

```
##      0      1
##  47839 205841
```

HvyAlcoholConsum, binary
- Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)

```
summary(health$HvyAlcoholConsum)
```

```
##      0      1
## 239424  14256
```

AnyHealthcare, binary
- Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?

```
summary(health$AnyHealthcare)
```

```
##      0      1
##  12417 241263
```

NoDocbcCost, binary
- Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?

```
summary(health$NoDocbcCost)
```

```
##      0      1
## 232326  21354
```

GenHlth, factor
- Would you say that in general your health is:
- 1 = Excellent
- 2 = Very Good
- 3 = Good
- 4 = Fair
- 5 = Poor

```
summary(health$GenHlth)
```

```
##     1     2     3     4     5
## 45299 89084 75646 31570 12081
```

MentHlth, numerical
- Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

```
summary(health$MentHlth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   3.185   2.000  30.000
```

PhysHlth, numerical
- Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

```
summary(health$PhysHlth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   4.242   3.000  30.000
```

DiffWalk, binary
- Do you have serious difficulty walking or climbing stairs?

```
summary(health$DiffWalk)
```

```
##      0      1
## 211005  42675
```

Sex, binary
- 0 = Female
- 1 = Male

```
summary(health$Sex)
```

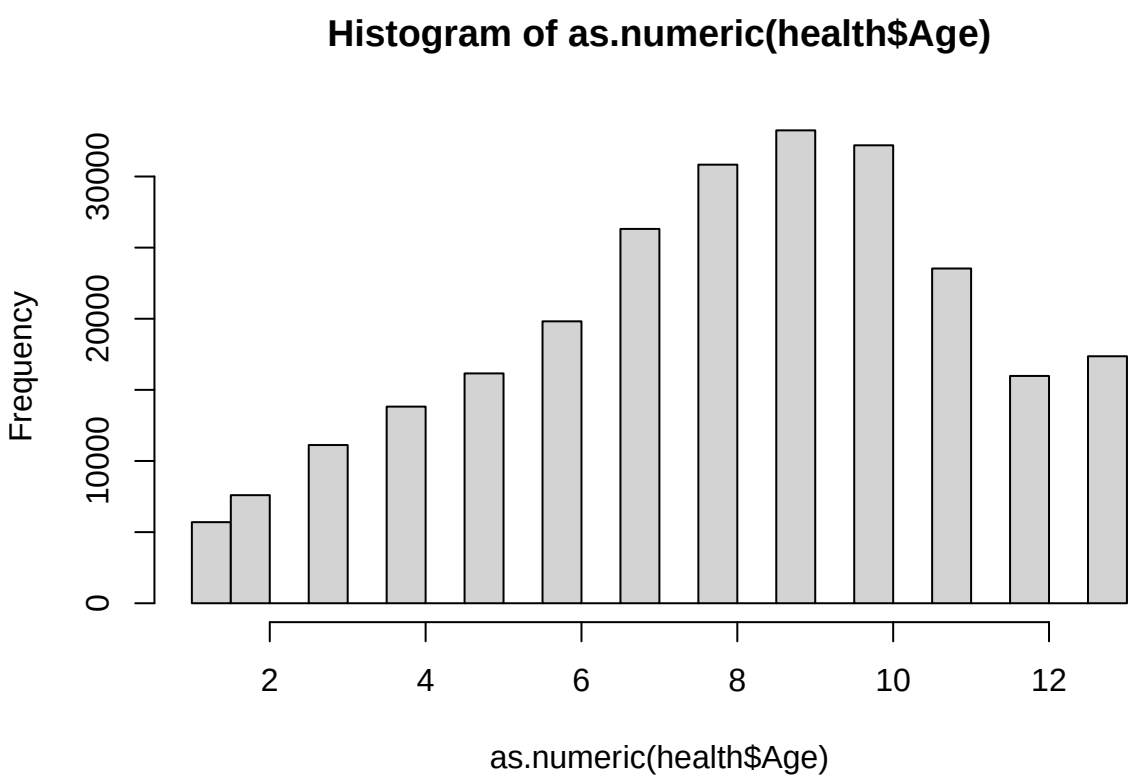
```
##      F      M
## 141974 111706
```

Age, factor
- 1 Age 18 to 24
- 2 Age 25 to 29
- 3 Age 30 to 34
- 4 Age 35 to 39
- 5 Age 40 to 44
- 6 Age 45 to 49
- 7 Age 50 to 54
- 8 Age 55 to 59
- 9 Age 60 to 64
- 10 Age 65 to 69
- 11 Age 70 to 74
- 12 Age 75 to 79
- 13 Age 80 or older
- 14 Don't Know / Refused to answer (I removed these as well)

```
summary(health$Age)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
##  5700  7598 11123 13823 16157 19819 26314 30832 33244 32194 23533 15980 17363
```

```
hist(as.numeric(health$Age))
```

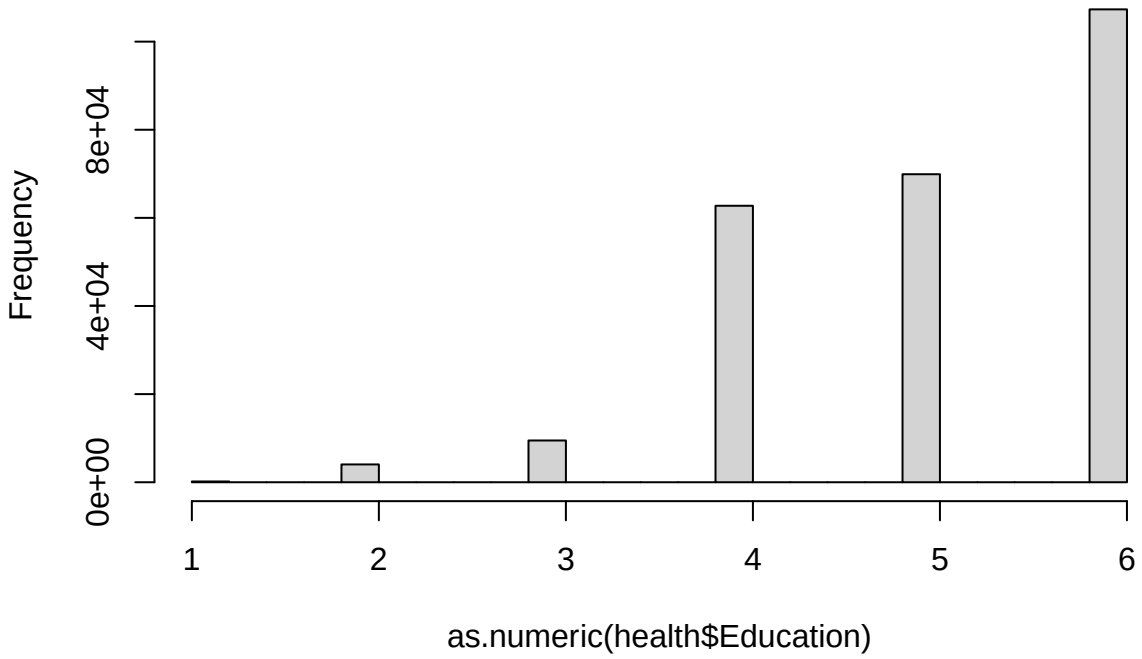


Education, factor
- What is the highest grade or year of school you completed?
- 1 Never attended school or only kindergarten
- 2 Grades 1 through 8 (Elementary)
- 3 Grades 9 through 11 (Some high school)
- 4 Grade 12 or GED (High school graduate)
- 5 College 1 year to 3 years (Some college or technical school)
- 6 College 4 years or more (College graduate)

```
summary(health$Education)
```

```
##      1      2      3      4      5      6
##    174   4043   9478  62750  69910 107325
```

```
hist(as.numeric(health$Education))
```

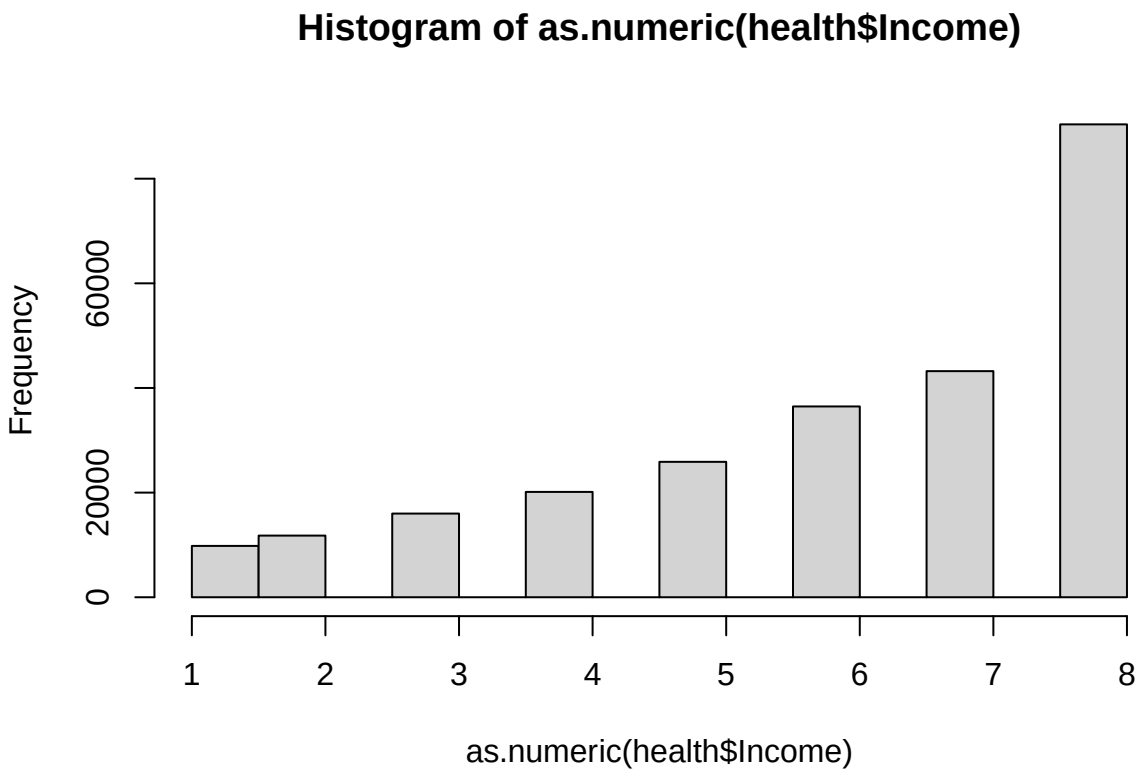


Income, factor
- Is your annual household income from all sources
- 1 Less than $10,000
- 2 Less than $15,000 ($10,000 to less than $15,000)
- 3 Less than $20,000 ($15,000 to less than $20,000)
- 4 Less than $25,000 ($20,000 to less than $25,000)
- 5 Less than $35,000 ($25,000 to less than $35,000)
- 6 Less than $50,000 ($35,000 to less than $50,000)
- 7 Less than $75,000 ($50,000 to less than $75,000)
- 8 $75,000 or more

```
summary(health$Income)
```

```
##     1     2     3     4     5     6     7     8
##  9811 11783 15994 20135 25883 36470 43219 90385
```

```
hist(as.numeric(health$Income))
```


Histogram of as.numeric(health$Income)
Frequency
0
20000
60000
1
2
3
4
5
6
7
8
as.numeric(health$Income)