Predicting Heart Attack

Jonathan Giguere, Jesse Borg, Ese Emuraye, Sarah Gates

Agenda

Introduction	Sarah
Data Cleaning	Jonathan
EDA	Jesse
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	Ese
Conclusion	Jesse

Agenda

Introduction	**Sarah**
Data Cleaning	Jonathan
EDA	Jesse
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	Ese
Conclusion	Jesse

Introduction

- Behavioral Risk Factor Surveillance System
  - Conducted by CDC in 50 states
  - Bi-annual telephone survey
  - Over 300 fields covering areas related to...
    - Education & income
    - Health information
    - Lifestyle habits
- Project focuses on factors related to heart attack
  - EDA section will explore relationships of a select number of fields to the binary occurence of heart attack
  - Modeling section will focus on a handful of variables that are the most related to heart attack

Data Source

- We obtained our data from the CDC website https://www.cdc.gov/brfss/annual_data/annual_2013.html as a .RData file.
    - 2013 data
    - 491,775 survey participants
    - 330 variables
- There are many categorical variables present in the dataset which will dictate which models we can create

Agenda

Introduction	Sarah
Data Cleaning	**Jonathan**
EDA	Jesse
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	Ese
Conclusion	Jesse

Data Cleaning and Feature Selection

- As a first step, we chose 45 of the 330 variables that we thought would be most helpful in predicting heart attacks.
    - We looked at other projects performed on BRFSS for guidance.
- Second, we renamed the variables to have names that are easier to work with.
- Next we checked for missing values and removed any variables with more than 100,000 missing entries.

Data Cleaning and Feature Selection

state	weight	month	sex
0	0	3	7
veteran	diabetes	high_bp	stroke
746	832	1420	1467
asthma	kidney_disease	health_coverage	gen_health
1559	1721	1904	1985
education_level	depression	heart_attack	employment_status
2274	2289	2587	3386
marital_status	angina	age5yr_bucket	sleep_time
3420	4423	4730	7387
mental_health	difficulty_walk	smokeless_tabac	smoke_100
8627	12764	14018	14920
alc_past_30	fruit_freq	exercise_30	green_veg_freq
19644	33798	34029	35157
income	told_high_chol	time_since_cholcheck	dr_visits_year
71426	71662	73178	154152
prediabetes	binge_alc	alc_perday_30	freq_smoke
257913	260444	260590	276983
bp_meds	aspirin_daily	soda_freq	pregnant
293201	355557	388580	414790
stop_smoke_year	has_asthma_now	work_hours_week	depressed_30
415421	426435	459413	491284
anxious_30			
491288			

Data Cleaning and Feature Selection

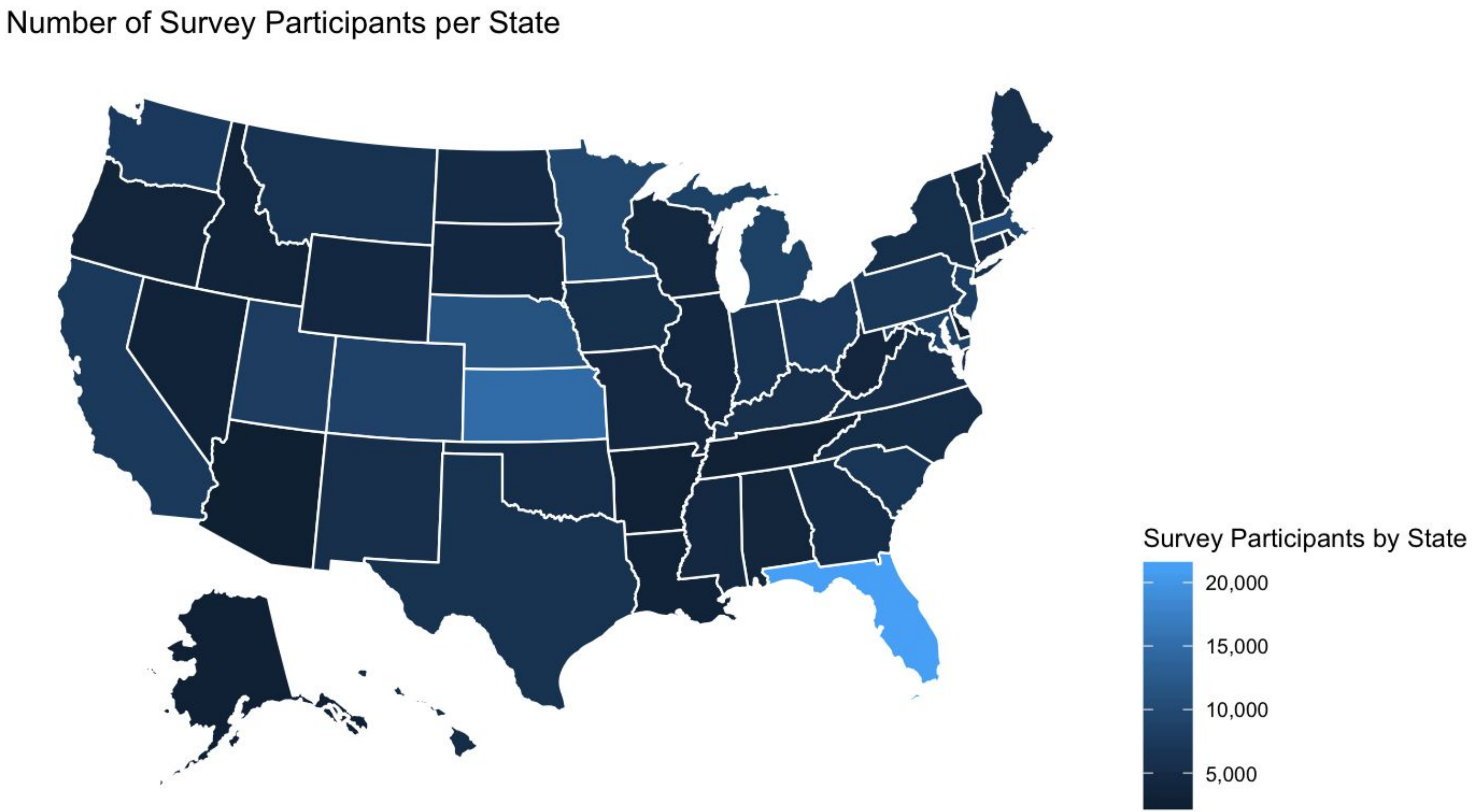
- To get complete data, we removed any rows with null values.
- Next, we changed factor levels to be 1s and 0s instead of Yes and No.
- Our final dataset has 313,247 records and 31 variables.

```
'data.frame':   313247 obs. of  31 variables:
 $ state              : Factor w/ 55 levels "0","Alabama",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ month              : Factor w/ 12 levels "January","February",..: 1 1 1 2 3 3 4 4 4 6 ...
 $ gen_health         : Factor w/ 5 levels "Excellent","Very good",..: 3 3 2 3 2 3 1 3 4 4 ...
 $ mental_health      : int  0 2 0 2 0 0 0 0 0 1 ...
 $ health_coverage    : Factor w/ 2 levels "1","0": 1 1 1 1 1 1 1 1 2 1 ...
 $ sleep_time         : int  6 9 8 6 8 6 8 8 3 8 ...
 $ high_bp            : Factor w/ 2 levels "1","0": 2 2 2 1 1 1 2 2 1 1 ...
 $ time_since_cholcheck: Factor w/ 4 levels "Within past year",..: 1 4 1 2 1 1 1 1 1 1 ...
 $ told_high_chol     : Factor w/ 2 levels "Yes","No": 2 2 1 2 1 1 1 2 1 1 ...
 $ heart_attack       : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 2 2 1 2 ...
 $ angina             : Factor w/ 2 levels "1","0": 2 2 2 2 2 1 2 2 2 2 ...
 $ stroke             : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 2 2 2 2 ...
 $ asthma             : Factor w/ 2 levels "1","0": 2 2 2 1 2 2 2 2 1 1 ...
 $ depression         : Factor w/ 2 levels "1","0": 1 1 2 2 2 2 2 2 1 1 ...
 $ kidney_disease     : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 2 2 2 1 ...
 $ diabetes           : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 2 2 2 1 ...
 $ veteran            : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 2 2 1 2 ...
 $ marital_status     : Factor w/ 6 levels "Married","Divorced",..: 1 1 1 1 2 3 1 1 1 2 ...
 $ education_level    : Factor w/ 6 levels "Never attended school or only kindergarten",..: 5 6 4 6 6 5 6 4 6 6 ...
 $ employment_status  : Factor w/ 7 levels "Employed","Self-employed",..: 1 1 6 6 1 6 6 4 3 6 ...
 $ income             : Factor w/ 8 levels "Less than $10,000",..: 8 8 7 6 8 6 8 4 1 8 ...
 $ weight             : Factor w/ 570 levels "",".b","100",..: 30 63 31 169 128 1 139 73 75 139 ...
 $ sex                : Factor w/ 2 levels "Male","Female": 2 2 2 1 2 2 1 2 1 2 ...
 $ difficulty_walk    : Factor w/ 2 levels "1","0": 2 1 2 2 2 1 2 2 1 2 ...
 $ smoke_100          : Factor w/ 2 levels "1","0": 2 1 2 1 2 1 2 2 1 1 ...
 $ smokeless_tabac    : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 1 2 2 2 ...
 $ alc_past_30        : int  0 220 208 210 0 202 101 0 0 0 ...
 $ fruit_freq         : int  301 203 306 302 206 320 101 202 215 101 ...
 $ green_veg_freq     : int  203 202 310 310 203 315 203 201 325 320 ...
 $ exercise_30        : Factor w/ 2 levels "1","0": 1 2 1 2 1 1 1 1 2 1 ...
 $ age5yr_bucket      : Factor w/ 13 levels "Age 18 to 24",..: 7 8 9 10 6 9 7 10 8 11 ...
```

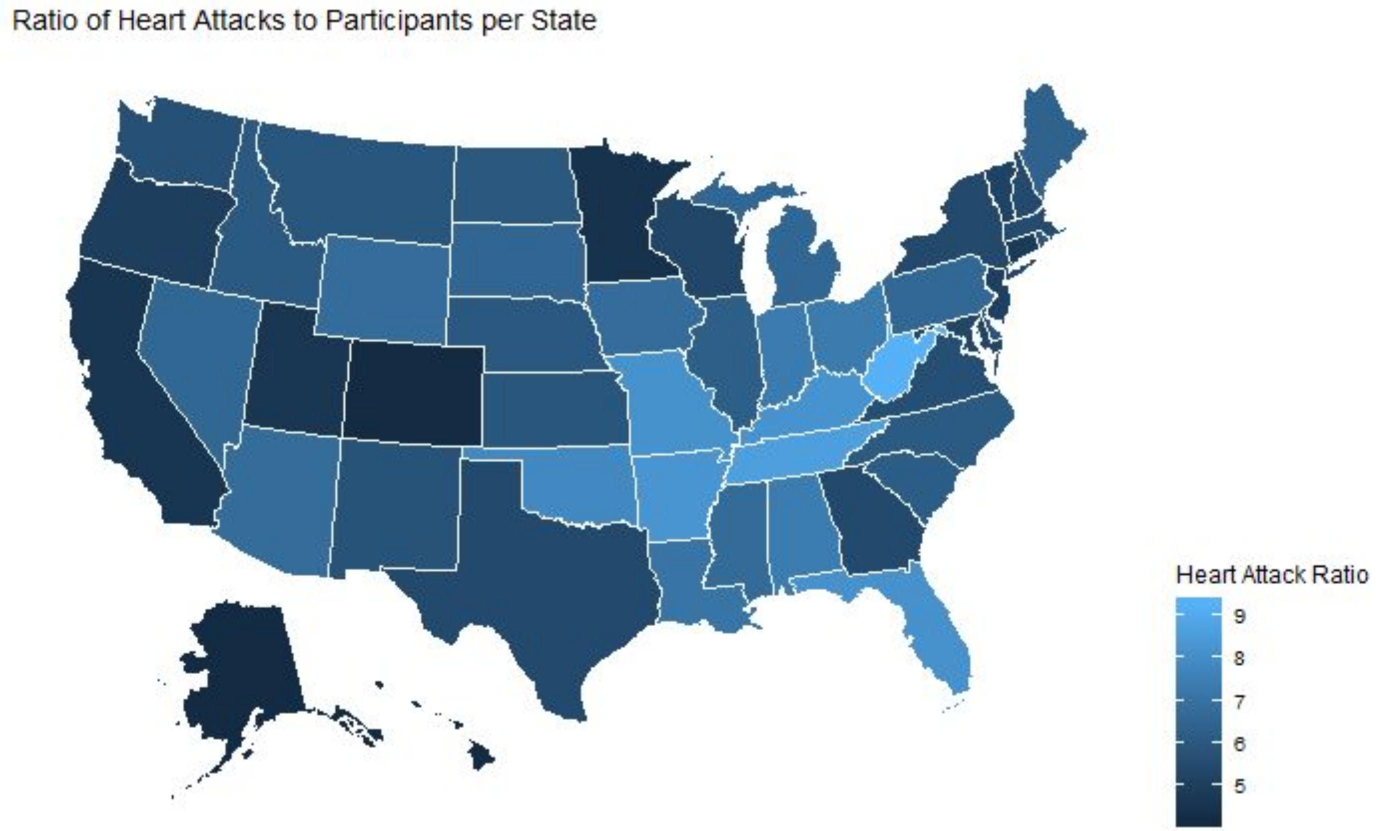
Agenda

Introduction	Sarah
Data Cleaning	Jonathan
EDA	**Jesse**
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	Ese
Conclusion	Jesse

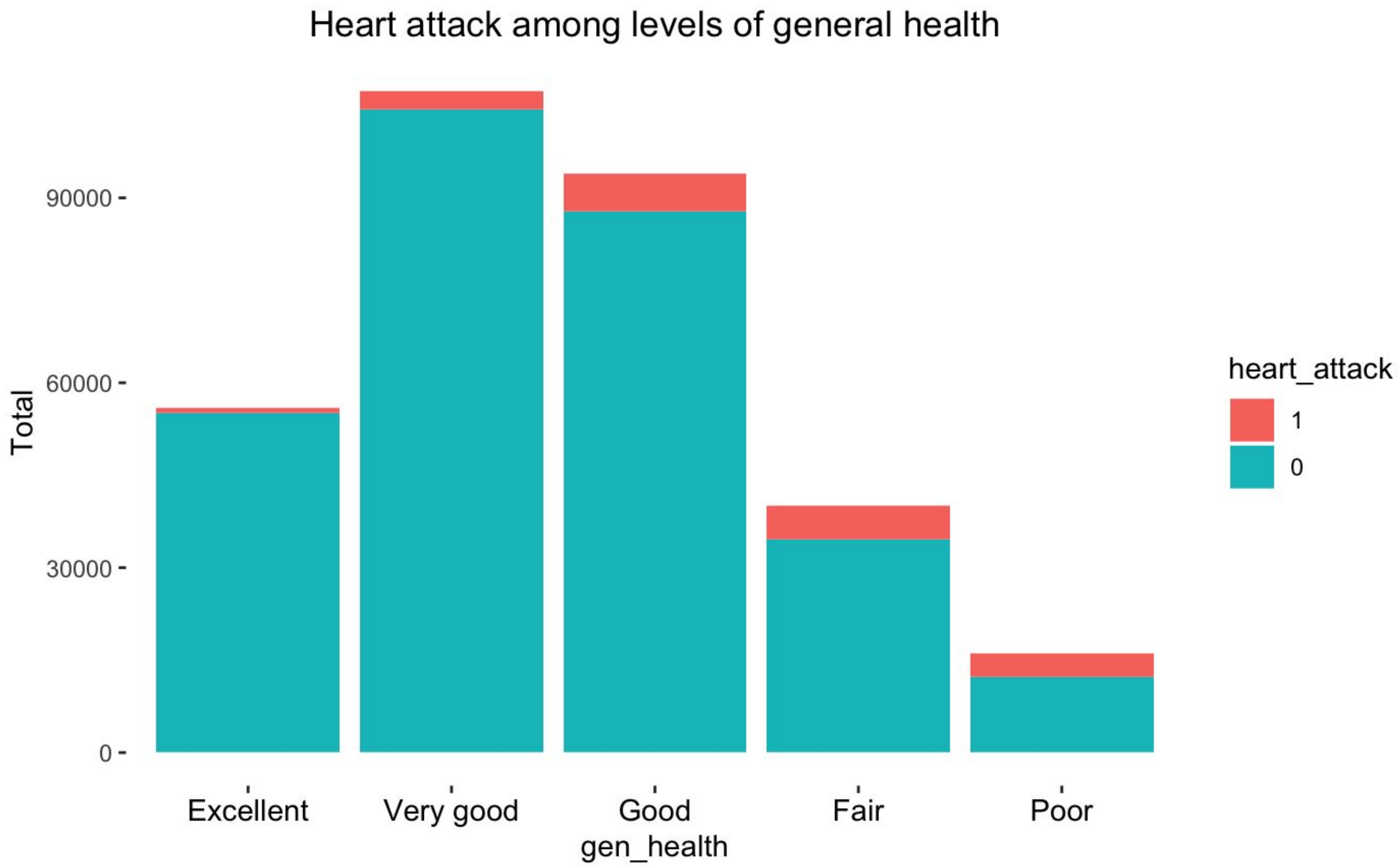
SMART Question - Which states have the most respondents?



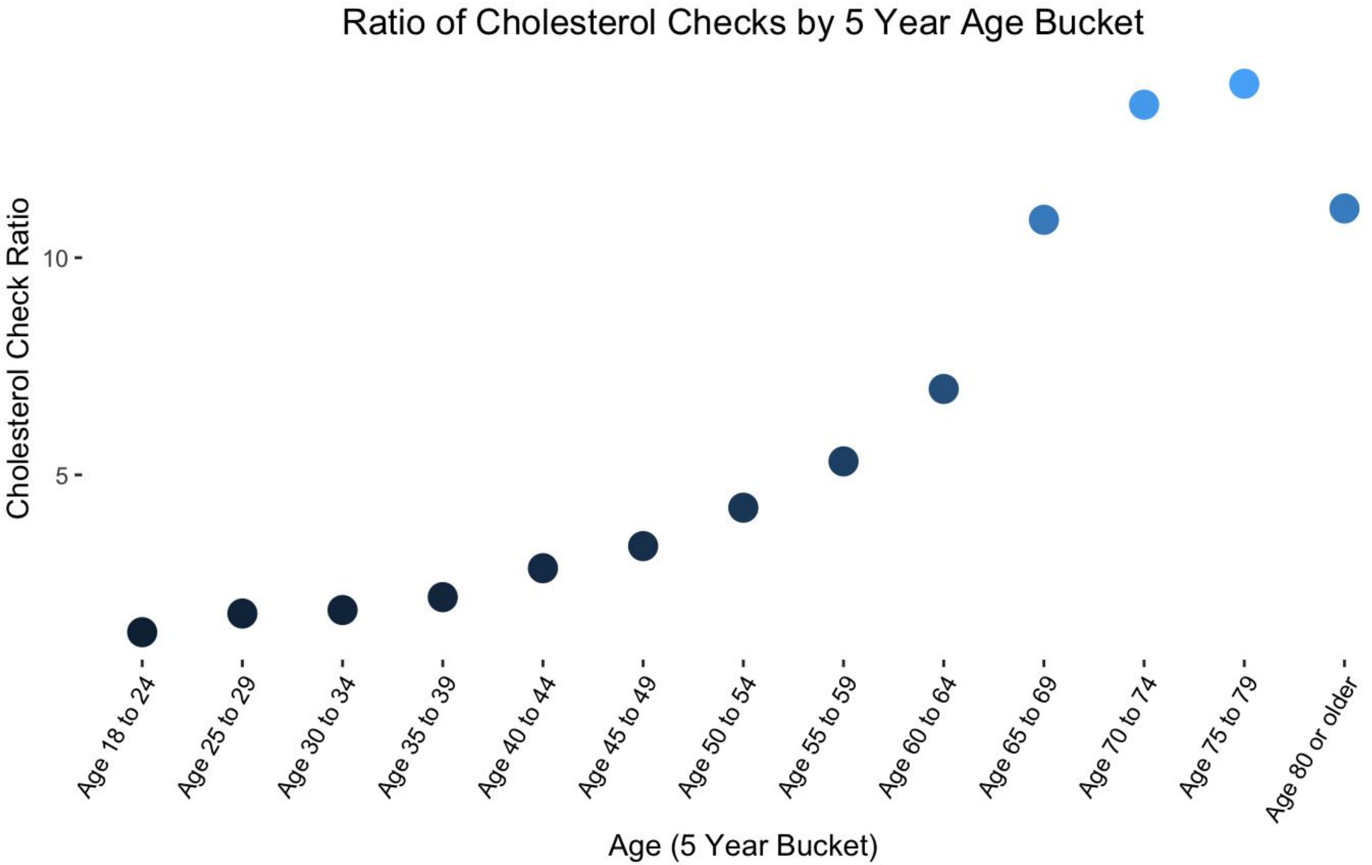
SMART Question - Which states have the most heart attacks?



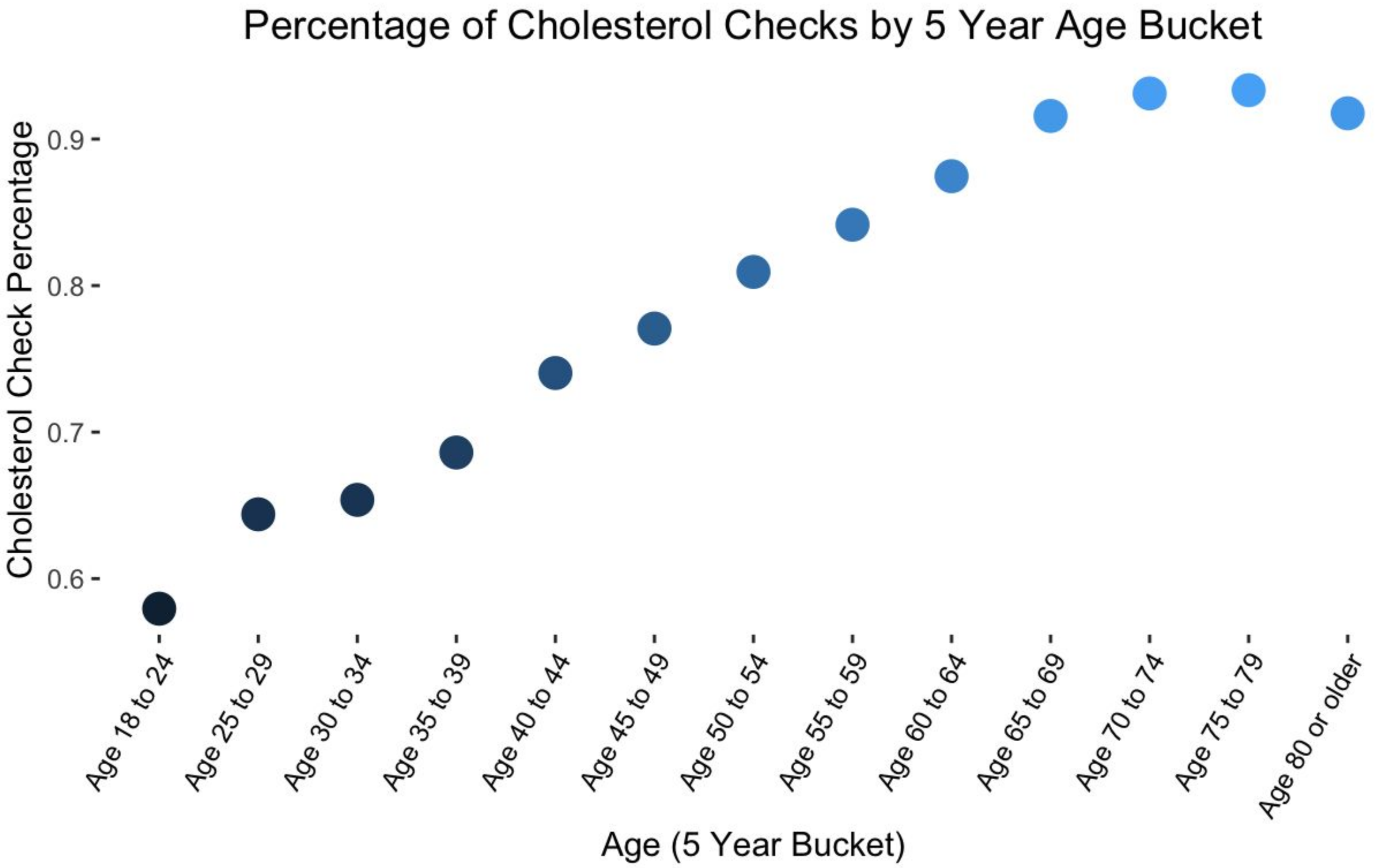
SMART Question - What is the relationship between heart attack and general health?



SMART Question - If a person is told they have high cholesterol, is there a pattern in their likeliness to check their cholesterol level frequently across age brackets?

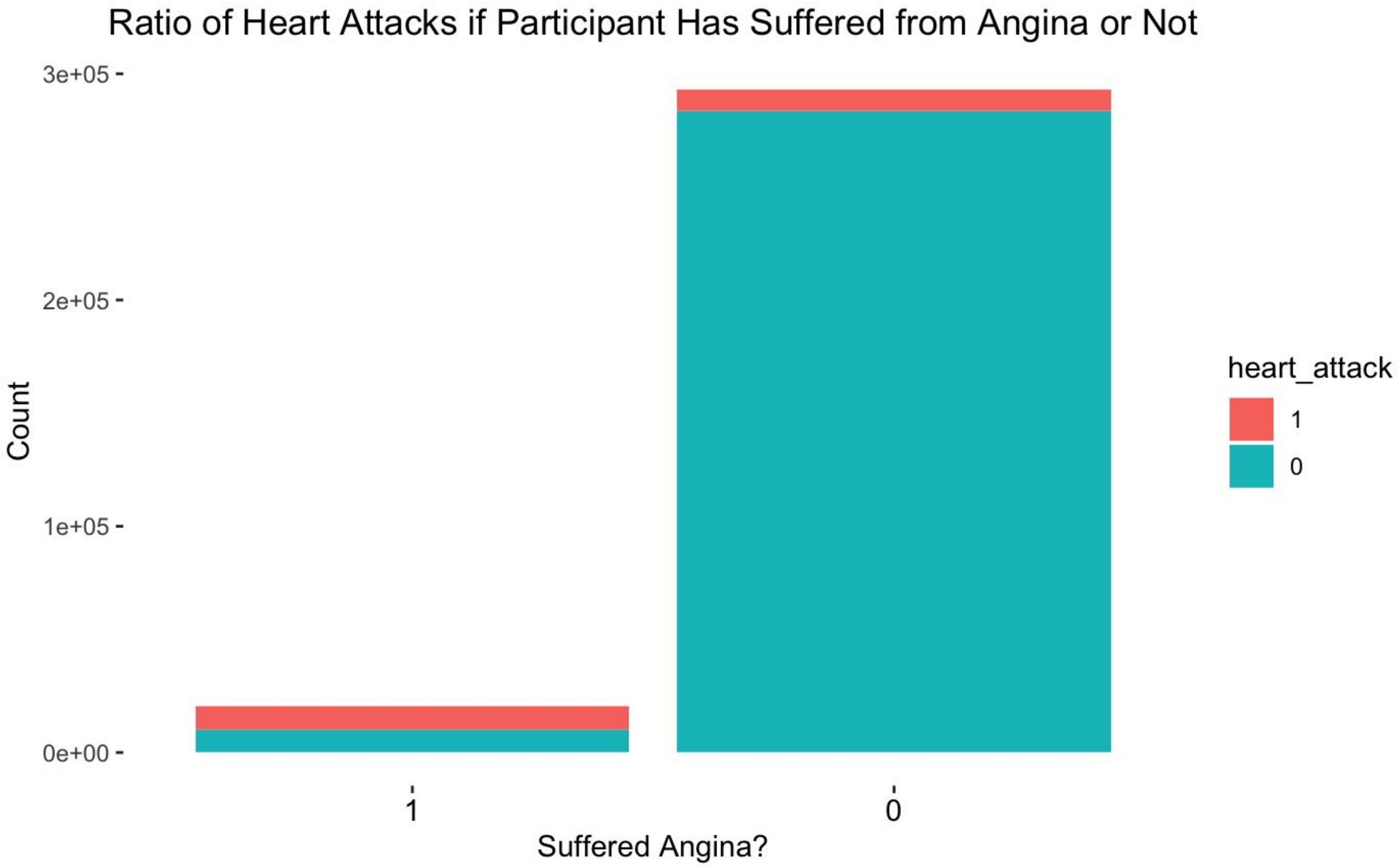


SMART Question - If a person is told they have high cholesterol, is there a pattern in their likeliness to check their cholesterol level frequently across age brackets?

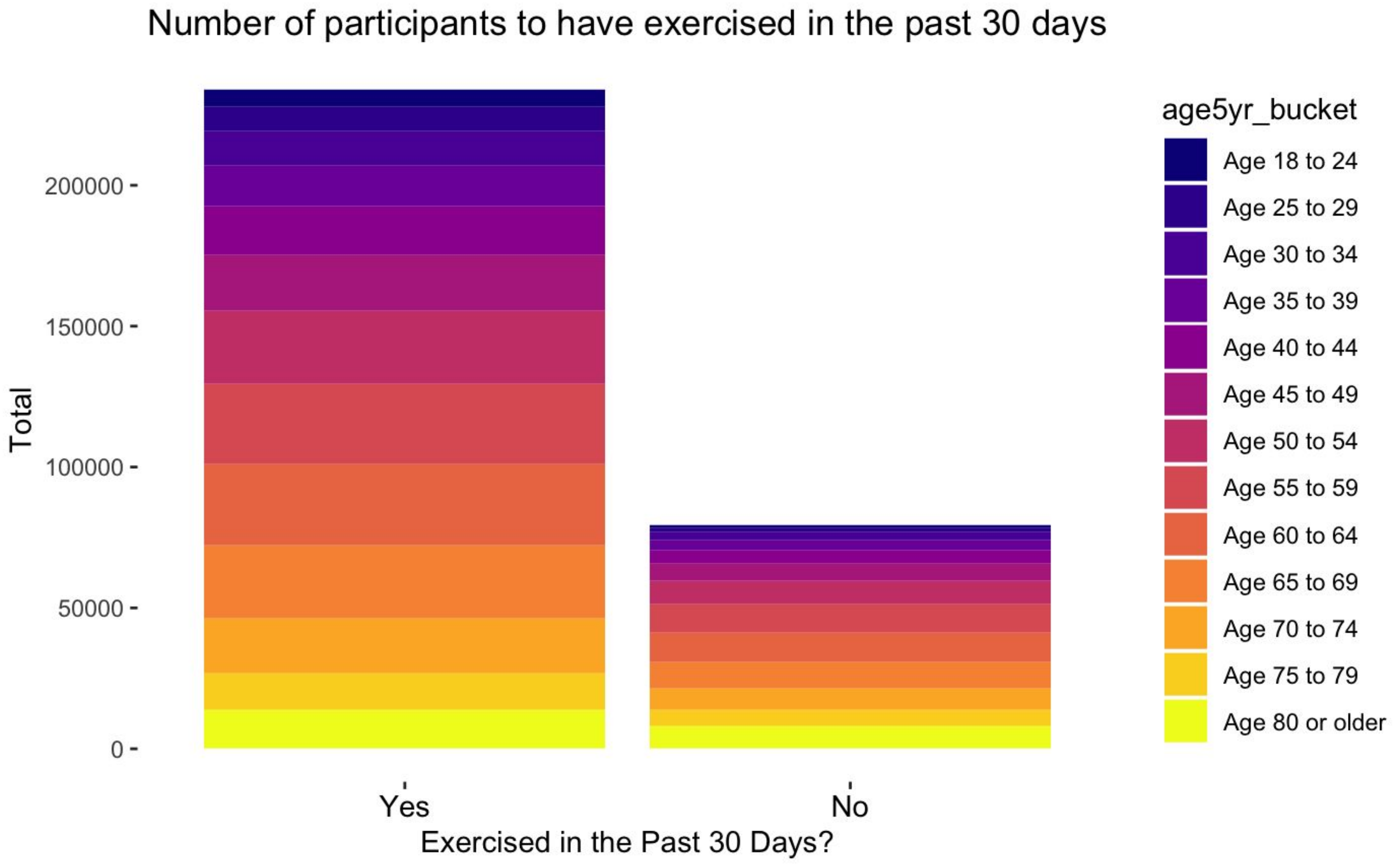


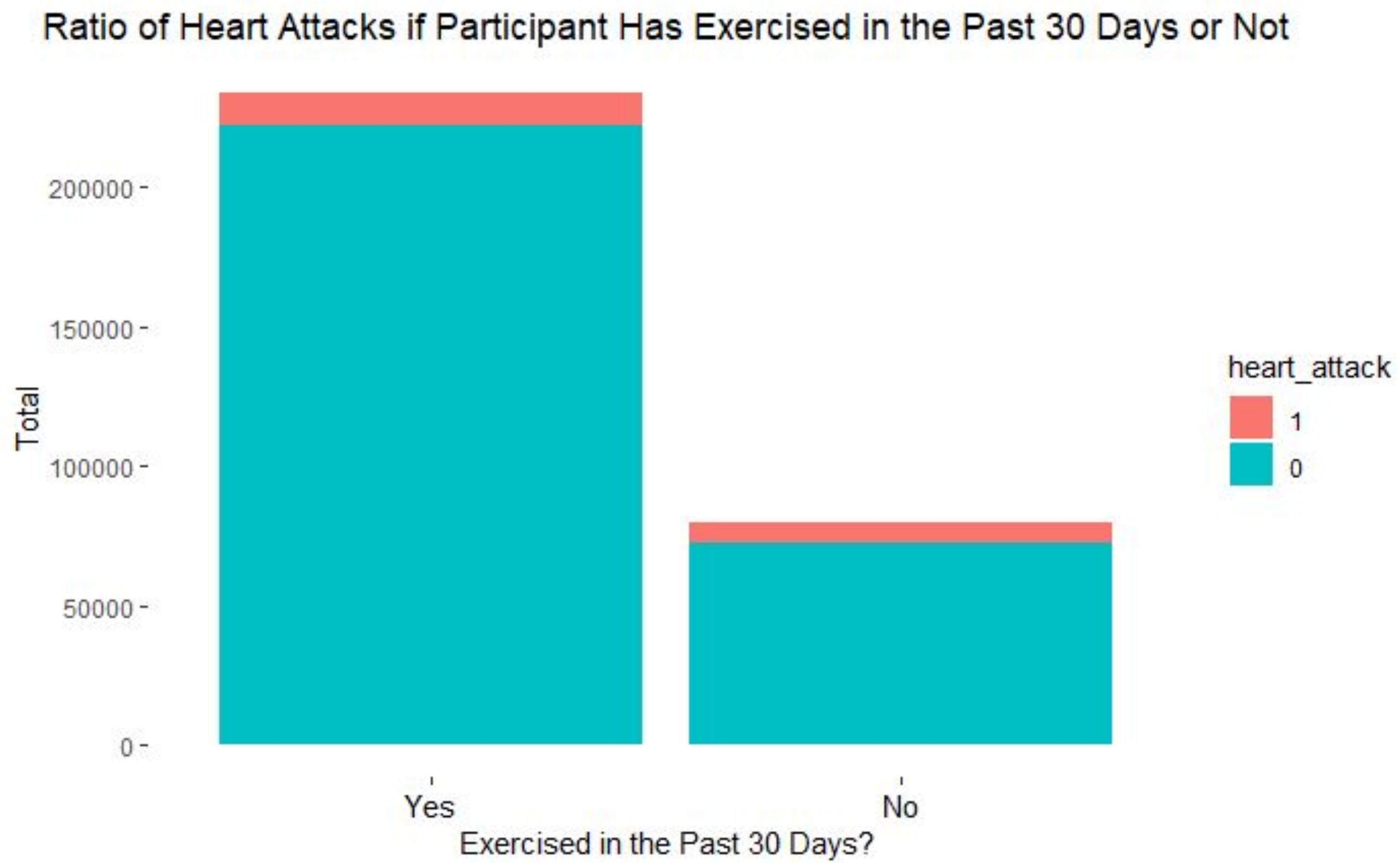
SMART Question - Is there a detectable pattern between heart attack and time since cholesterol check across age brackets?

SMART Question - What is the relationship between angina and heart attacks?

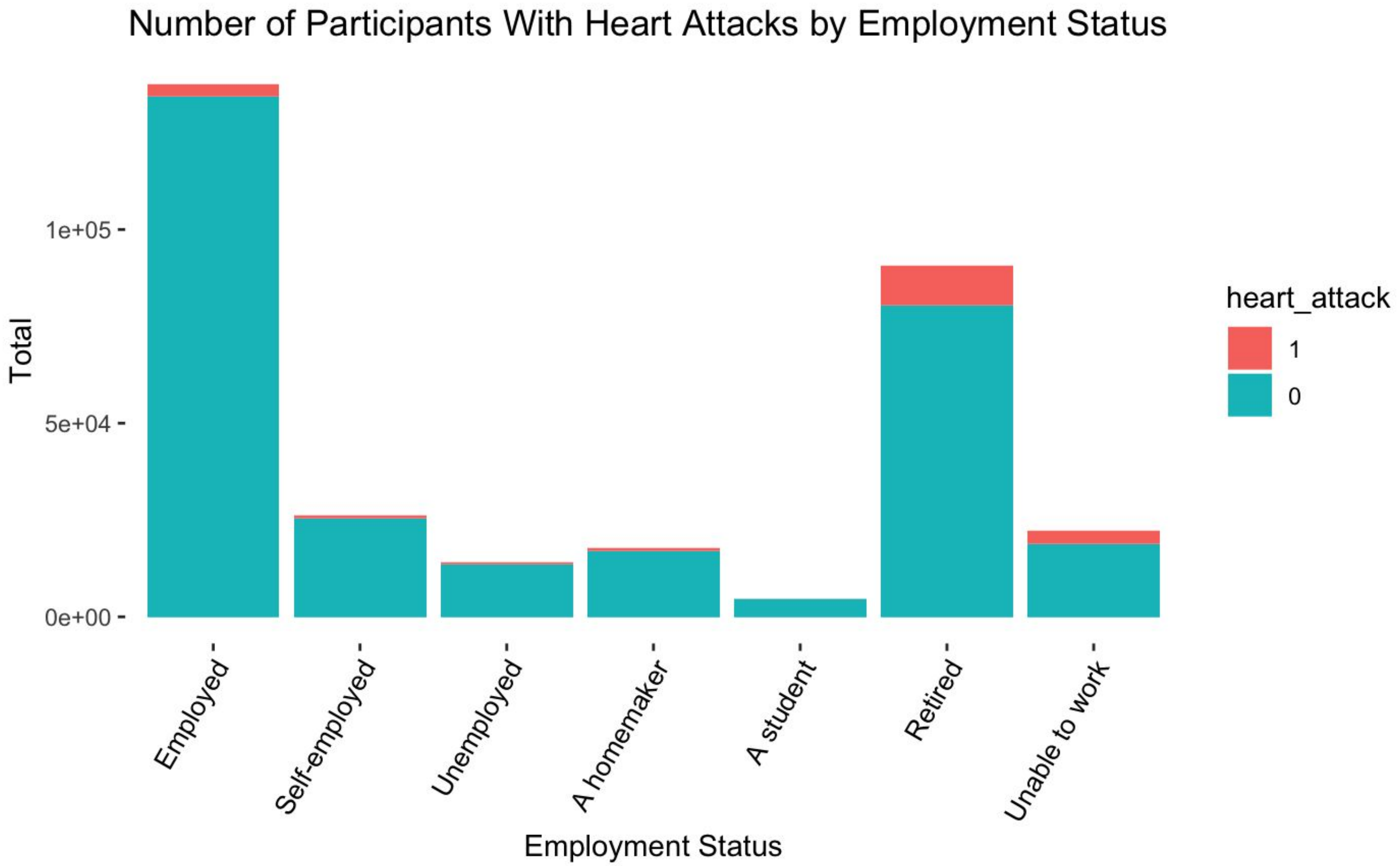


SMART Question - What effect does age have on exercise?

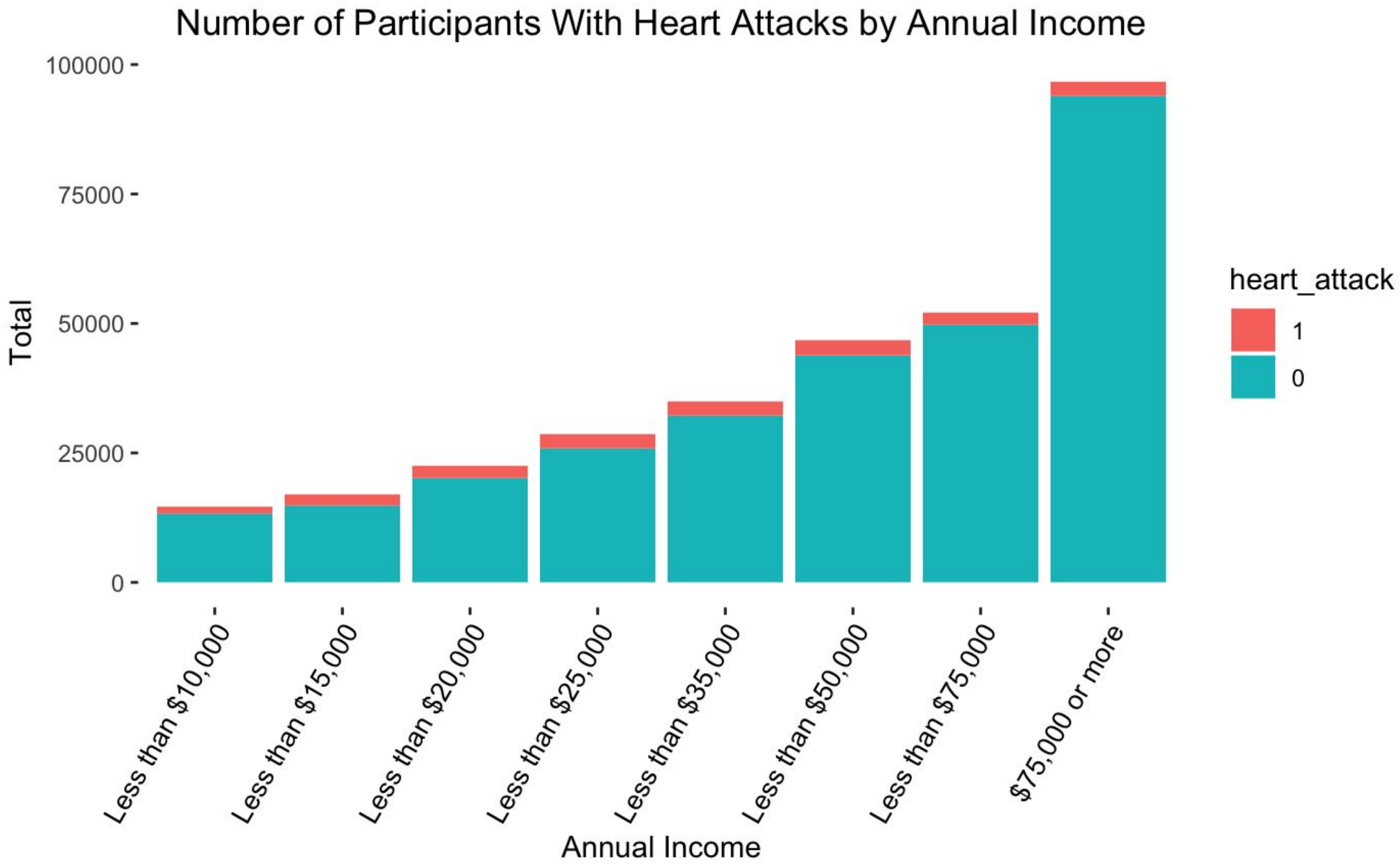



Ratio of Heart Attacks if Participant Has Exercised in the Past 30 Days or Not
Total
200000
150000
100000
50000
0
Yes
No
Exercised in the Past 30 Days?
heart_attack
1
0

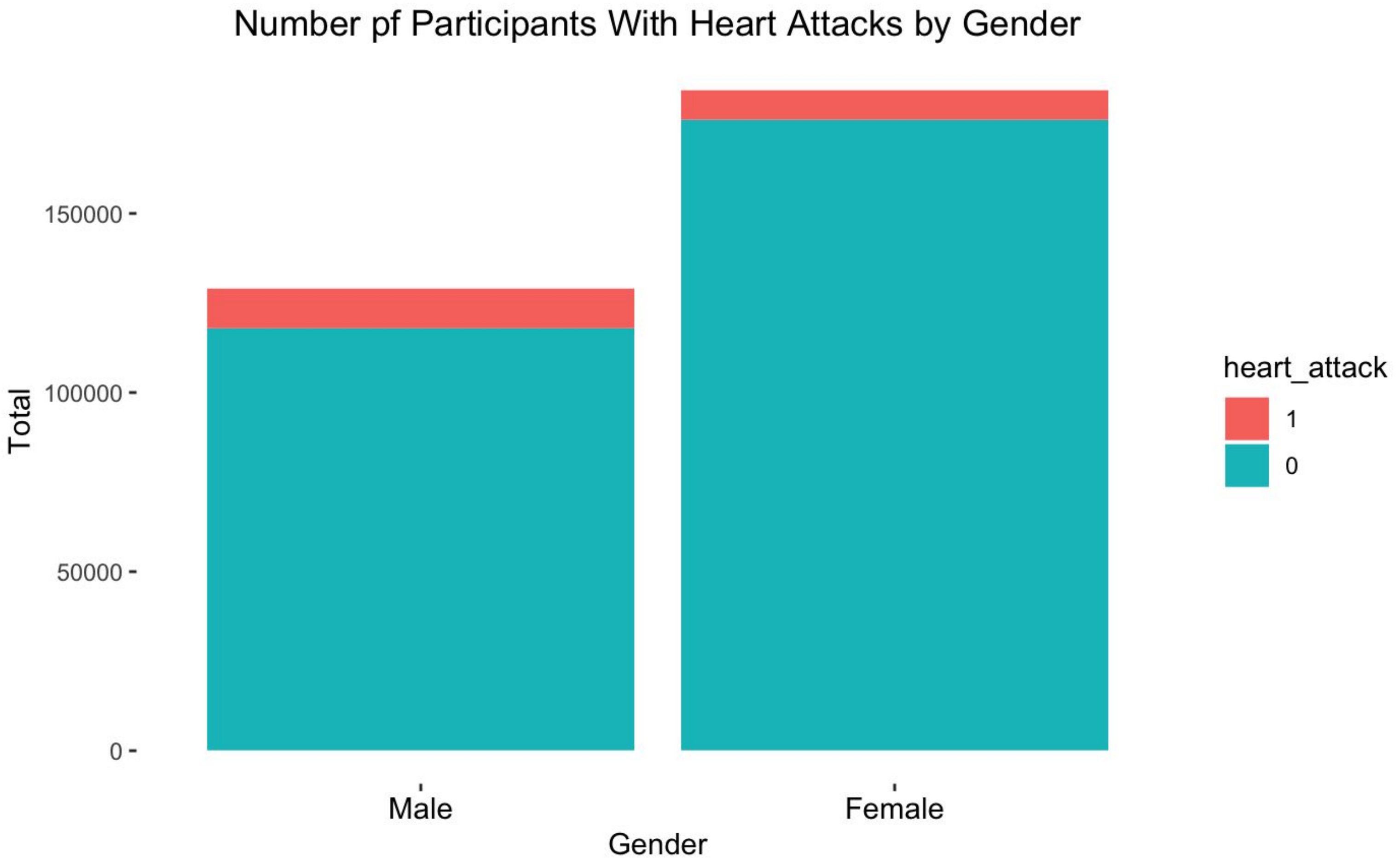
SMART Question - What is the relationship of heart attacks and employment status?



SMART Question - What is the relationship between heart attack and annual income levels?



SMART Question - Do men or women have more heart attacks?



Agenda

Introduction	Sarah
Data Cleaning	Jonathan
EDA	Jesse
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	Ese
Conclusion	Jesse

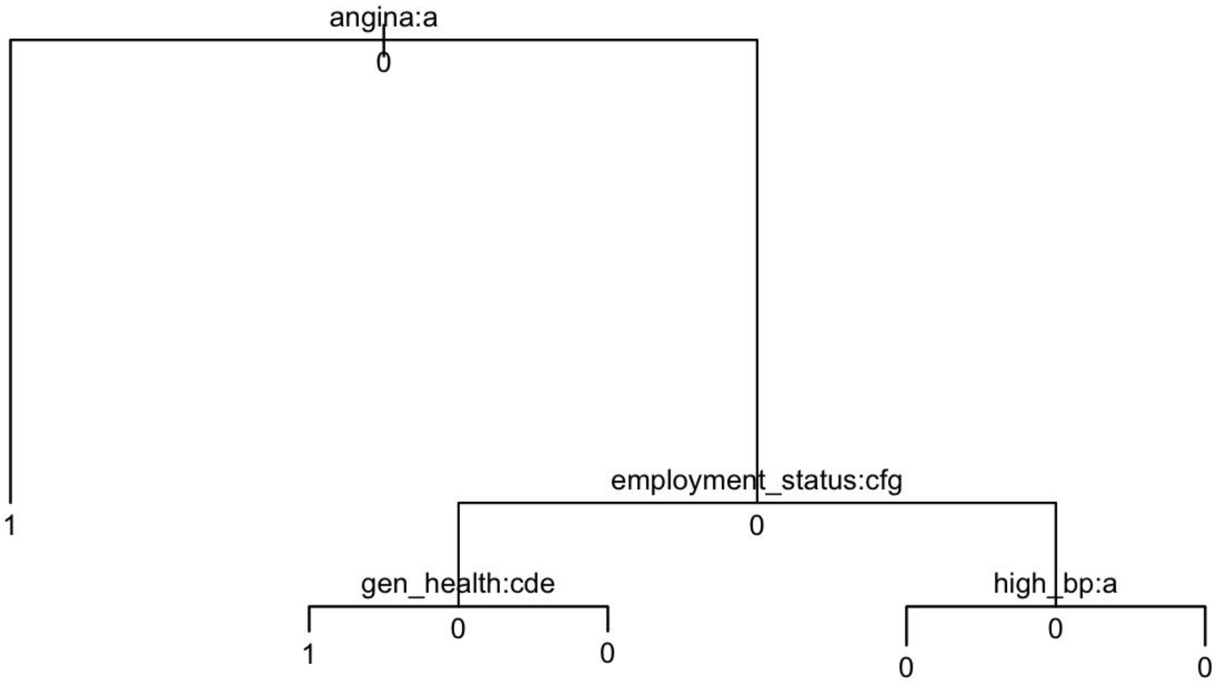
Hypothesis Tests for Feature Selection

Categorical Variable	Chi-Test P-value < 0.05?	Include?
gen_health	TRUE	Yes
mental_health	TRUE	Yes
health_coverage	TRUE	Yes
high_bp	TRUE	Yes
time_since_cholcheck	TRUE	Yes
told_high_chol	TRUE	Yes
angina	TRUE	Yes
stroke	TRUE	Yes
ashtma	TRUE	Yes
depression	TRUE	Yes
kidney_disease	TRUE	Yes
diabetes	TRUE	Yes
veteran	TRUE	Yes
marital_status	TRUE	Yes
education_level	TRUE	Yes
employment_status	TRUE	Yes
income	TRUE	Yes
sex	TRUE	Yes
difficulty_walk	TRUE	Yes
smoke_100	TRUE	Yes
smokeless_tabac	FALSE	No
exercise_30	TRUE	Yes
age5yr_bucket	TRUE	Yes

Model Building Considerations

- A lot of categorical predictor variables so have to make a classification tree first.
- We will then perform logistic regression on the predictor variables identified as most important in our classification tree to see which performs better.
- When creating our first model, we realized that our dataset was imbalanced with only 6% of respondents having heart attacks.
  - To balance the dataset, we subset all respondents with heart attack and then randomly sampled the same number of records from the respondents without heart attack.
  - After balancing, we randomly split training (70%) and test datasets (30%).

Classification Tree

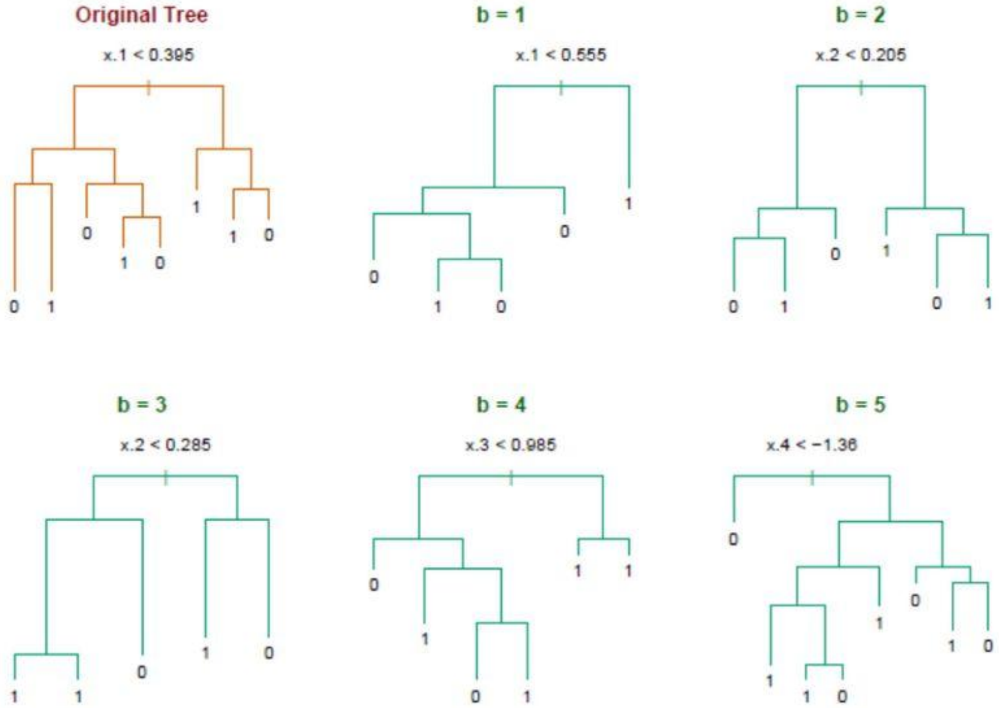


Classification Tree - Evaluation and Pruning

```
tree.pred    1    0
        1 4502 1322
        0 1258 4434
[1] 0.7759639
```

```
[1] 5 4 2 1
[1]  6088  6088  6660 13558
```

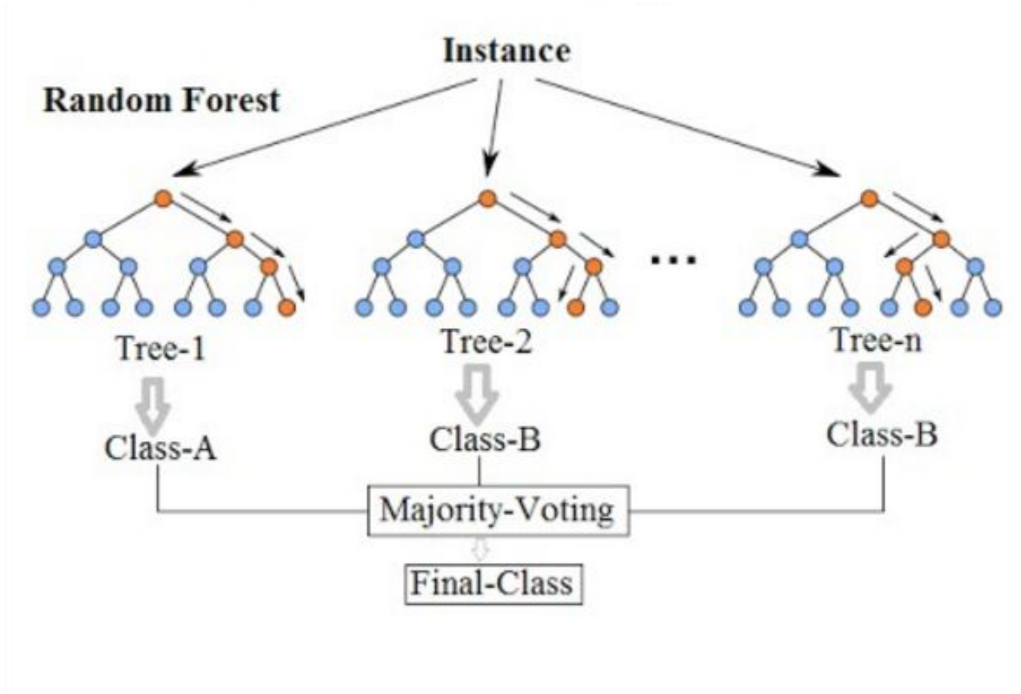
Bagged Tree



Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

```
tree.pred_bagged    1    0
               1 4466 1076
               0 1294 4680
[1] 0.7941994
```

Random Forest




```
tree.pred_randForest    1    0
                   1 4668 1185
                   0 1092 4571
[1] 0.8022751
```

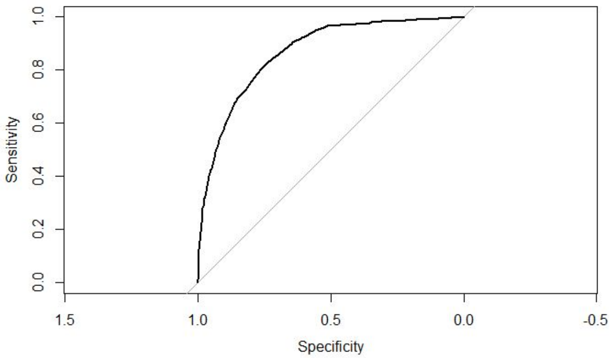
Logistic Regression

- Based on the findings from tree models, the following variables were used to create a logistic regression model:
  - Angina
  - Employment Status
  - General Health
  - High Blood Pressure
- All are factor variables, will this be able to create a strong model?

Logistic Regression

```
Call:
glm(formula = heart_attack ~ angina + employment_status + gen_health +
    high_bp, family = "binomial", data = bal_hrt_attack_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4129  -0.6417   0.2637   0.7430   2.8386

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -0.98883    0.07761 -12.742  < 2e-16 ***
angina0                         2.86465    0.05119  55.962  < 2e-16 ***
employment_statusSelf-employed -0.45763    0.06697  -6.833 8.30e-12 ***
employment_statusUnemployed    -0.51834    0.07695  -6.736 1.63e-11 ***
employment_statusA homemaker   -0.24288    0.07711  -3.150  0.00163 **
employment_statusA student      0.71730    0.23391   3.067  0.00217 **
employment_statusRetired       -1.08726    0.03909 -27.814  < 2e-16 ***
employment_statusUnable to work -0.83801   0.06013 -13.936  < 2e-16 ***
gen_healthVery good            -0.48681    0.06292  -7.737 1.02e-14 ***
gen_healthGood                 -1.03022    0.06124 -16.822  < 2e-16 ***
gen_healthFair                 -1.51037    0.06622 -22.809  < 2e-16 ***
gen_healthPoor                 -1.93484    0.08049 -24.039  < 2e-16 ***
high_bp0                        0.74870    0.03352  22.335  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37302  on 26907  degrees of freedom
Residual deviance: 24241  on 26895  degrees of freedom
AIC: 24267

Number of Fisher Scoring iterations: 5
```

- Model Evaluation
  - Growth/decay factors
  - CIs of each factor level
  - Hosmer Lemeshow
- ROC
  - Applied model to test data
  - AUC = .8661



Agenda

Introduction	Sarah
Data Cleaning	Jonathan
EDA	Jesse
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	**Ese**
Conclusion	Jesse

Model Evaluation

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- Metrics for evaluation for binary classification
  - Accuracy: fraction of heart attack predictions our model got right

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  - Precision: fraction of positive identifications (heart attack predictions) actually correct

$$\text{Precision} = \frac{TP}{TP + FP}$$

  - Recall: fraction of actual positives (heart attack predictions) identified correctly

$$\text{Recall} = \frac{TP}{TP + FN}$$

Model Evaluation

- Metrics for evaluation for binary classification
  - F1 Score : An harmonic mean of precision and recall

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Model Evaluation

Statistic	Logistic Regression	Classification Tree	Bagged Tree	Random Forrest
Accuracy	0.780653	0.7759639	0.79281	0.8013199
Specificity	0.7532936	0.7703266	0.8104587	0.7916956
Sensitivity	0.8145914	0.7815972	0.7751736	0.8109375
Precision	0.7269097	0.7730082	0.8036357	0.7957411
Recall	0.8145914	0.7815972	0.7751736	0.8109375
F1 Score	0.7682569	0.777279	0.7891481	0.8032674

Model Evaluation

Statistic	Logistic Regression	Classification Tree	Bagged Tree	Random Forrest
Accuracy	0.780653	0.7759639	0.79281	0.8013199
Specificity	0.7532936	0.7703266	0.8104587	0.7916956
Sensitivity	0.8145914	0.7815972	0.7751736	0.8109375
Precision	0.7269097	0.7730082	0.8036357	0.7957411
Recall	0.8145914	0.7815972	0.7751736	0.8109375
F1 Score	0.7682569	0.777279	0.7891481	0.8032674

Agenda

Introduction	Sarah
Data Cleaning	Jonathan
EDA	Jesse
Model Building	
Classification Tree	Jonathan
Bagged Tree	Jonathan
Random Forest	Ese
Logistic Regression	Sarah
Model Evaluation	Ese
Conclusion	**Jesse**

Conclusions

Of the 4 predictive models:

- The Random Forest model gave the best model performance with the highest accuracy value
- The Logistic Regression model is preferred for initial screening for heart attack because it had the highest sensitivity and, therefore, detection rate.
- We confirmed previously reported risk factors and also identified agina, general health, employment status, and high blood pressure as potential risk factors related to heart attack, with angina being the significantly most important predictor

Thank you!

BRFSS™

Behavioral Risk Factor Surveillance System

Questions?