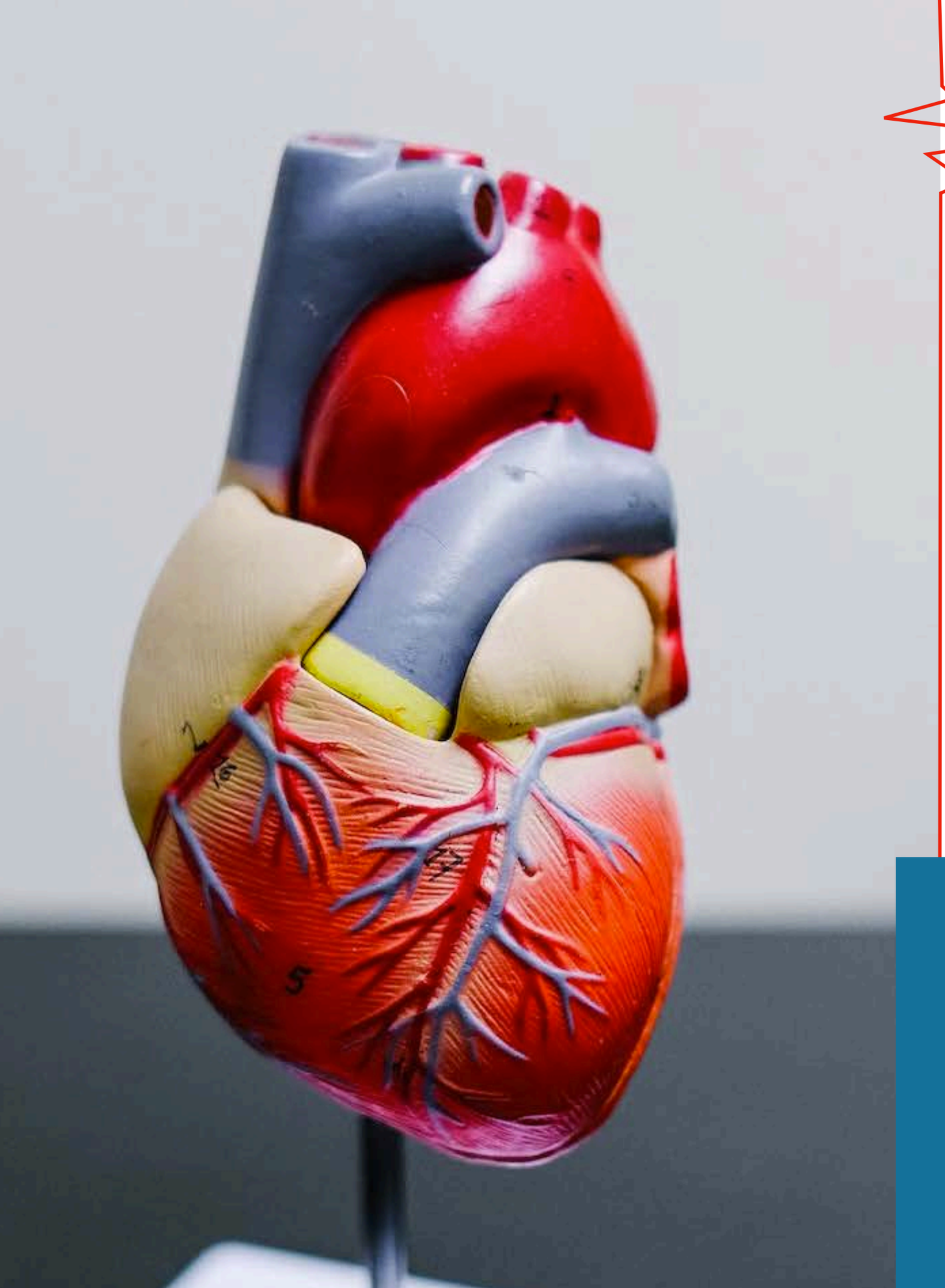


FINAL PROJECT

PROJECT PRESENTATION

AFFENDI ABDULLAH

INTRODUCTION

The America International Hospital is running a campaign to better educate the public on the awareness of heart diseases amongst its citizen.

They aim is to better prepare potential Heart Disease patients to plan for the probable illness and take good methodical measures as a prevention plan

INITIAL RESEARCH INTO THE DATABASE

The hospital had engaged us to help research into their dataset and provide key points to better execute their campaign

They also wish to have a system to which they are able to predict new patients with similar lifestyle and engage them for dialogue session with their representatives

MAIN OBJECTIVE

PREDICT IF A PERSON WITH SIMILAR LIFESTYLE CHOICES BE PRONE TO A HEART DISEASE

1. Find the total number of people getting Heart Disease in the original data

2. What is the number of man and woman in the original data

3. What are the ratios of men and women that has a history of heart disease?

4. Which race has the highest count of having a heart disease

5. For patients with Heart Disease, does their preexisting condition contribute to their ailment?

6. At which age category that that these Heart Disease cases seemed to be happen to?

INSIGHTS

INITIAL RESEARCH INTO THE DATABASE

Before we dive into the questions, we equipped ourselves with the following fact from the datasheet

DATA SHEET DESCRIPTION

INITIAL RESEARCH INTO THE DATABASE

ROWS , COLUMNS

319795, 18

COLUMN TITLES

HeartDisease	2 —— Yes & No Values
BMI	3604 —— Float
Smoking	2 —— Yes & No Values
AlcoholDrinking	2 —— Yes & No Values
Stroke	2 —— Yes & No Values
PhysicalHealth	31 —— Float
MentalHealth	31 —— Yes & No Values
DiffWalking	2 —— Yes & No Values
Sex	2 —— Male Female

COLUMN TITLES

Sex	2 —— Yes & No Values
AgeCategory	13 —— String
Race	6 —— Categorical Value
Diabetic	4 —— Yes & No + Outlier
PhysicalActivity	2 —— Yes & No Values
GenHealth	5 —— Numerical Values
SleepTime	24 —— Numerical Value
Asthma	2 —— Yes & No Values
KidneyDisease	2 —— Yes & No Values
SkinCancer	2 —— Yes & No Values

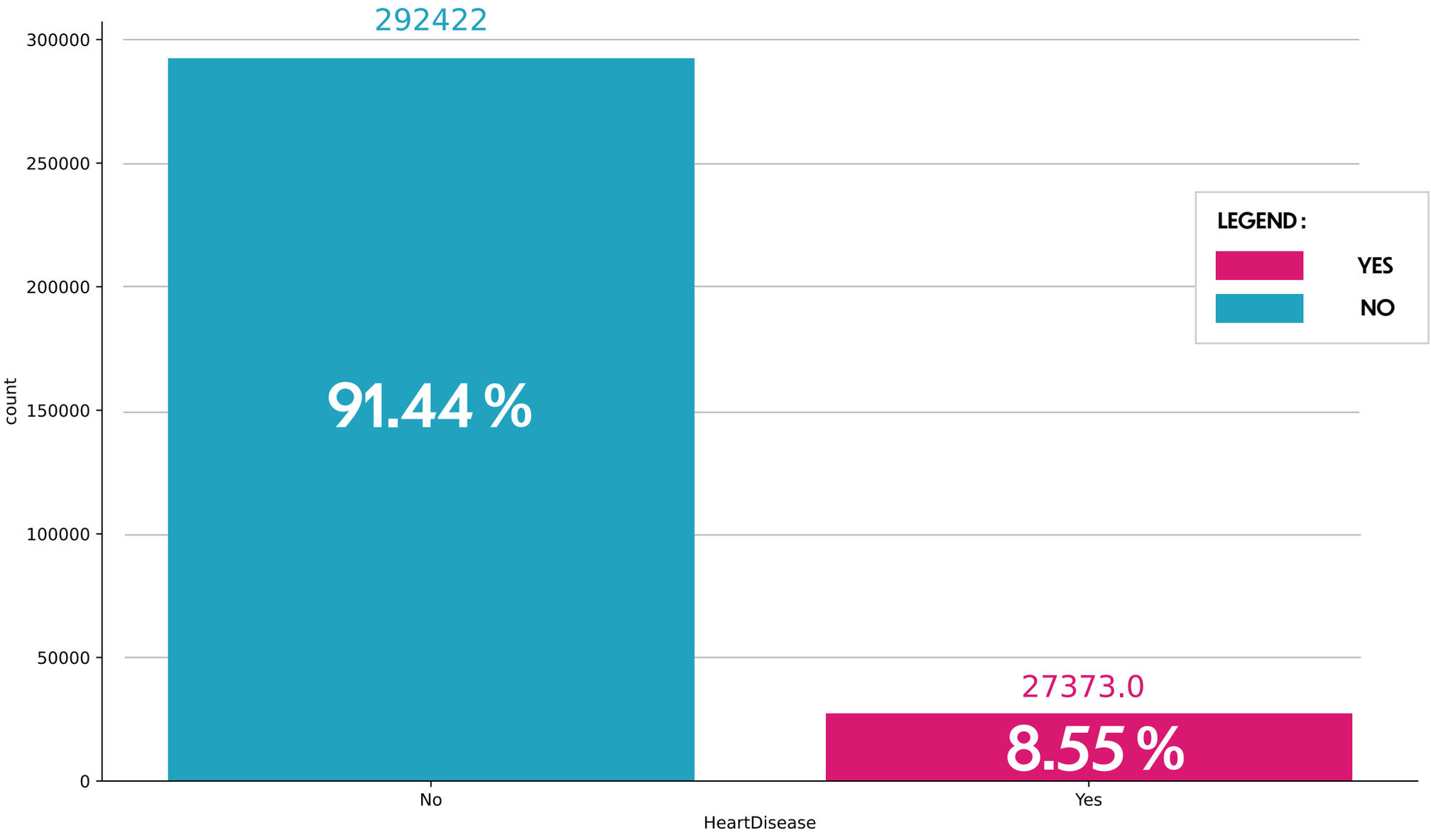
QUESTION 1

ORIGINAL RAW DATA

FIND THE TOTAL NUMBER OF PEOPLE GETTING HEART DISEASE IN THE ORIGINAL DATA

QUESTION 1

The number of people in the data that has heart disease



CONCLUSION

The number of Negative cases greatly outweighs the number of Positive case in the dataset. This dataset is highly imbalance

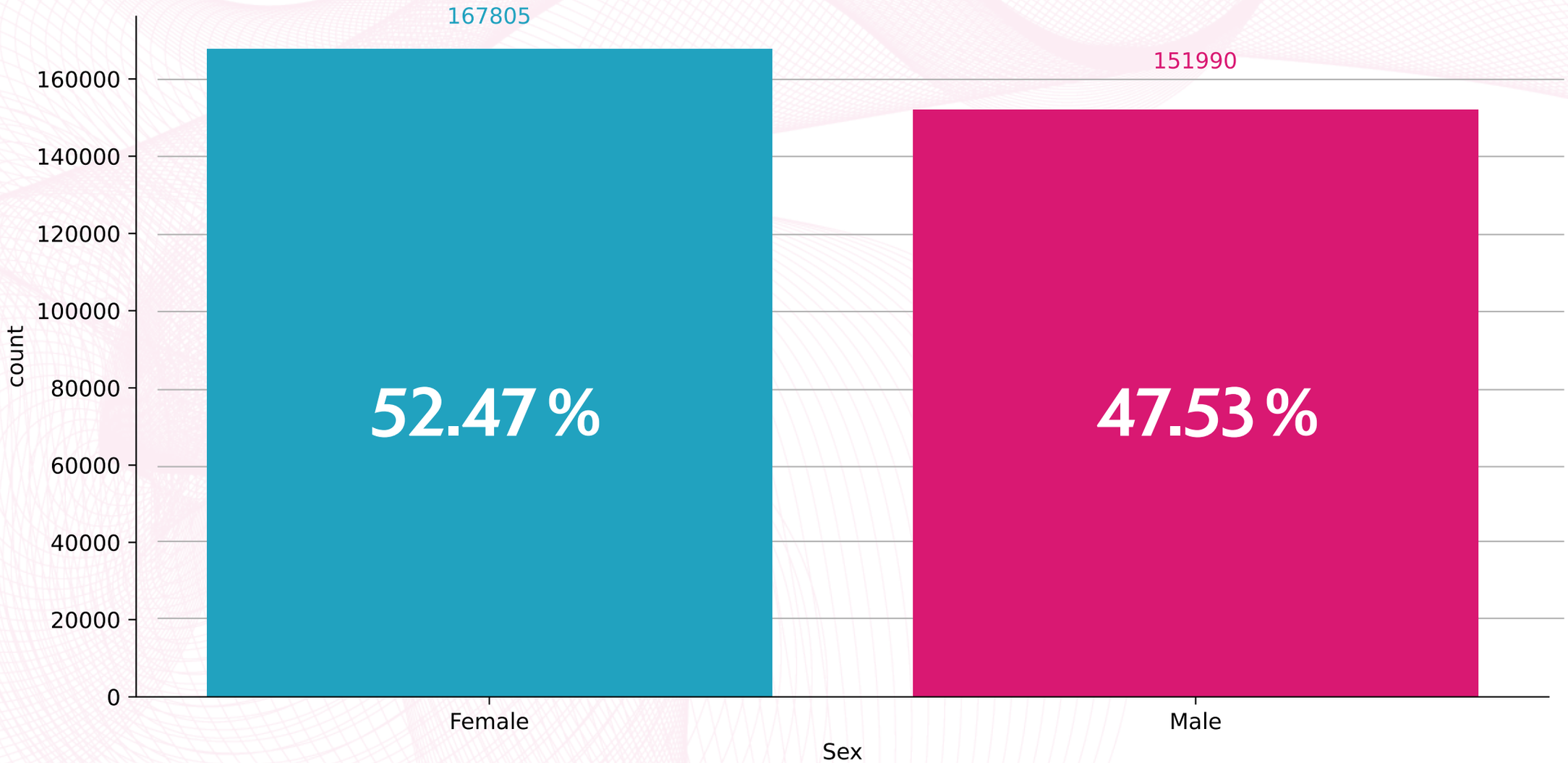
QUESTION 2

ORIGINAL RAW DATA

WHAT ARE THE RATIOS OF MEN AND WOMAN THAT IS IN THE ORIGINAL DATAFRAME?

QUESTION 2

What are the ratios of men and woman that is in the original data frame



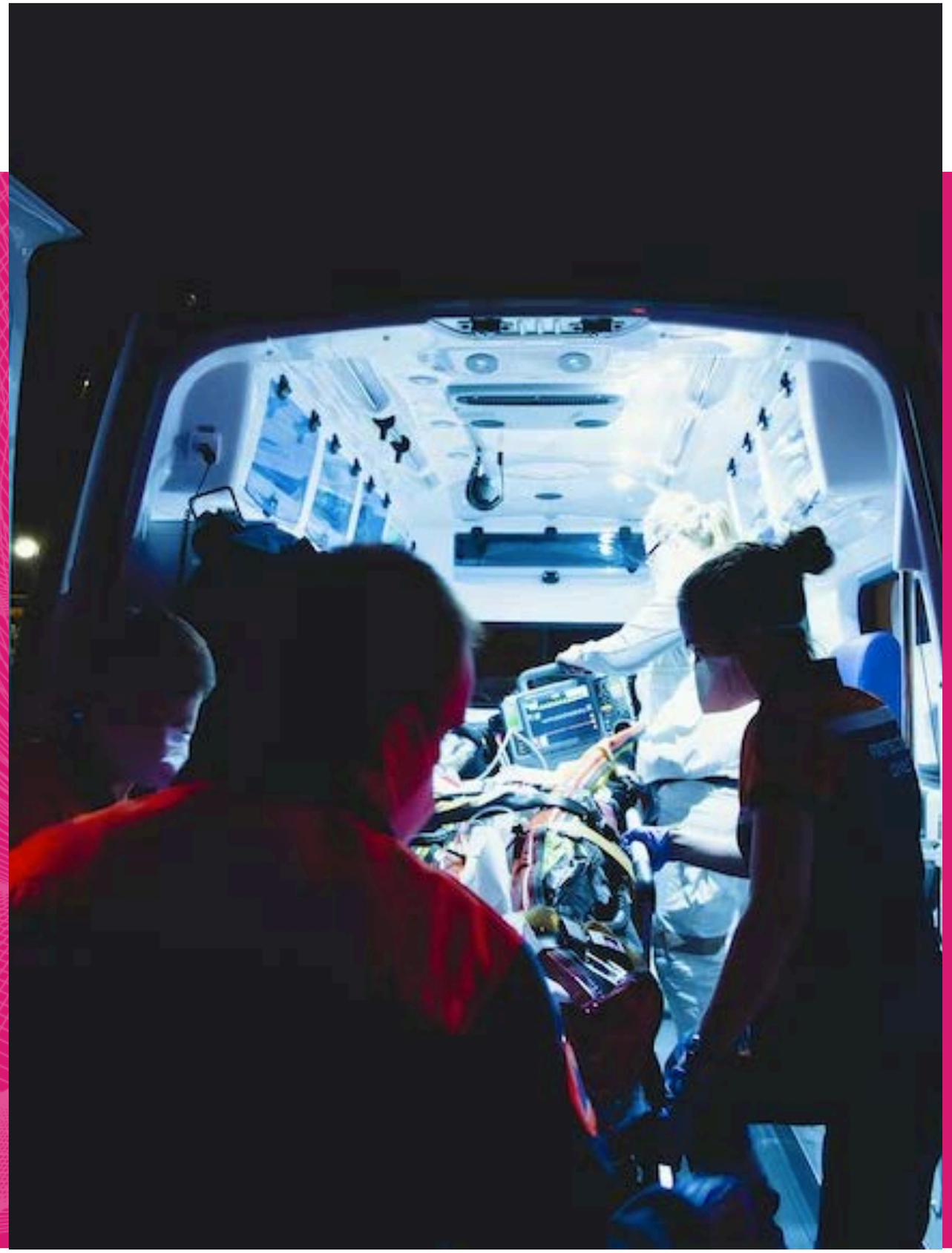
CONCLUSION

There are slightly more women than men that is present in the original data.

QUESTION 3

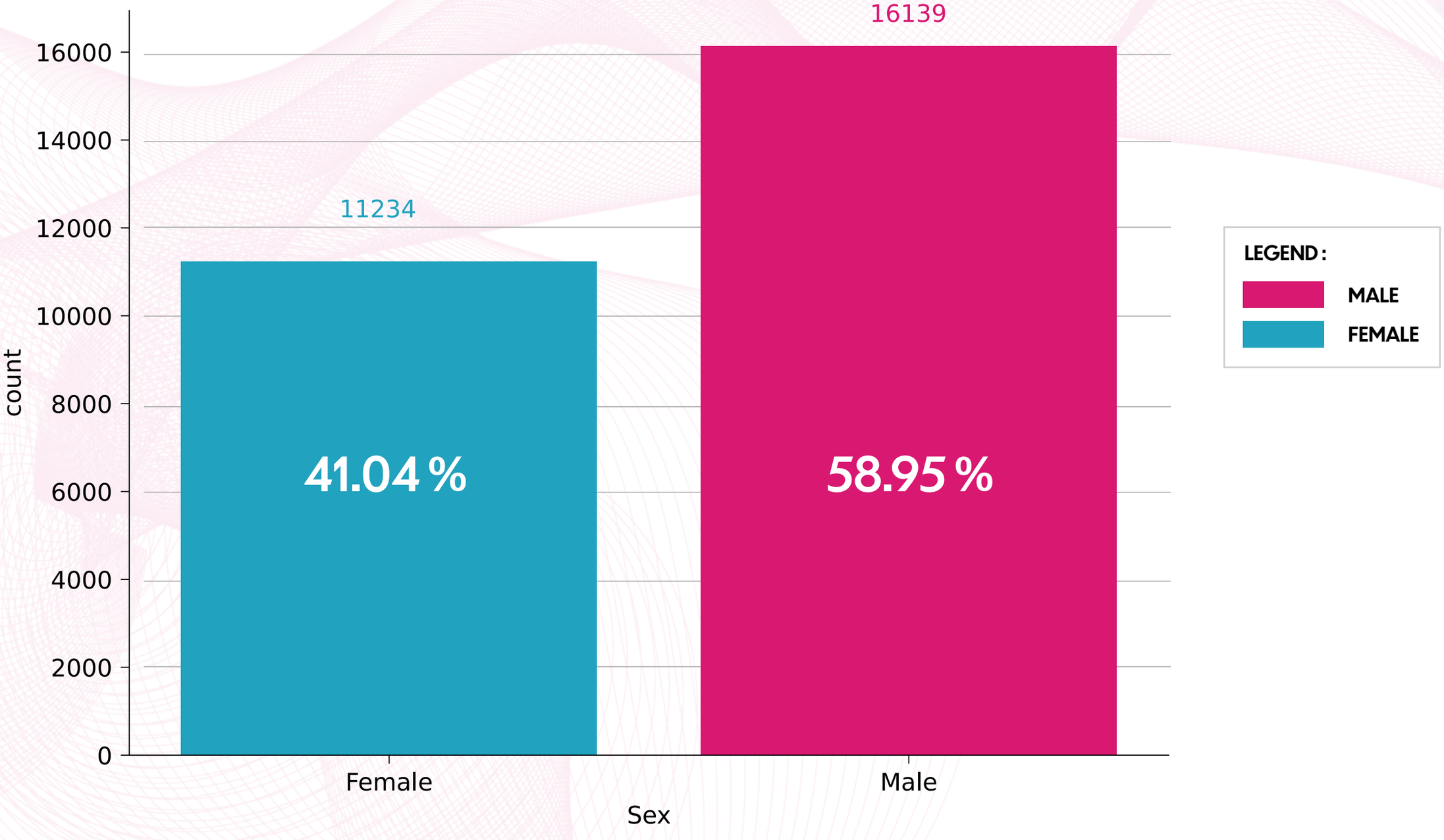
HEART DISEASE DATA

WHAT ARE THE RATIOS OF MEN AND WOMEN THAT HAS A HISTORY OF HEART DISEASE?



QUESTION 3

What are the ratios of men and women that has a history of heart disease?



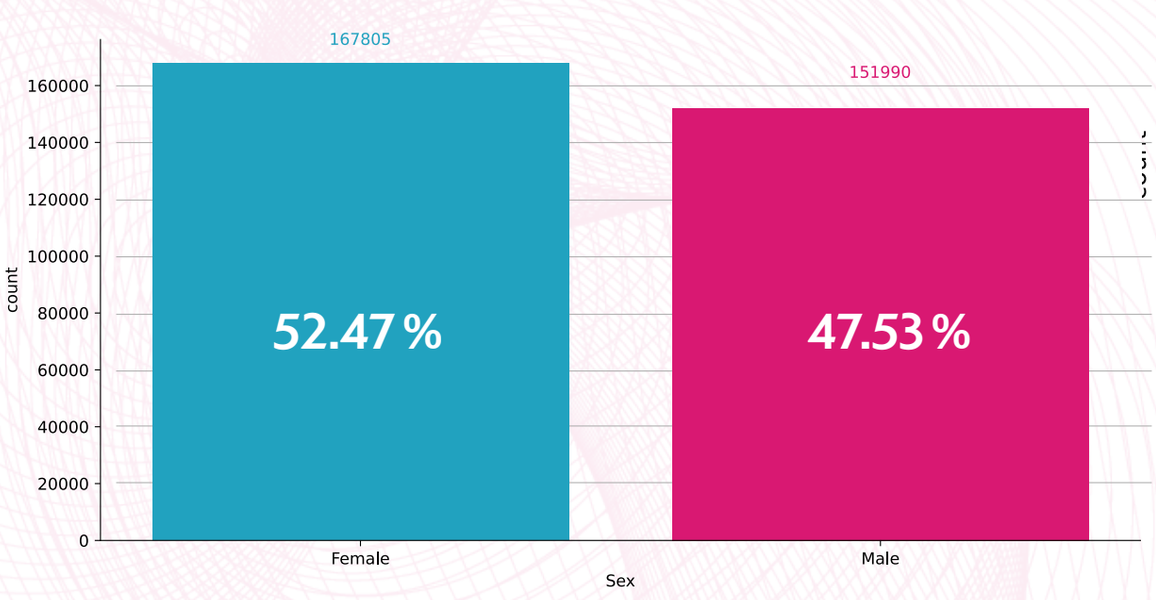
CONCLUSION

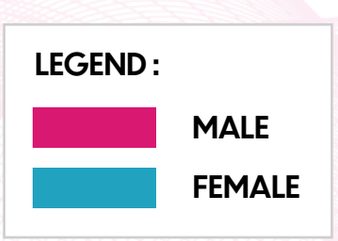
The majority of those that has a history of heart disease are men while women are closely behind

QUESTION 3

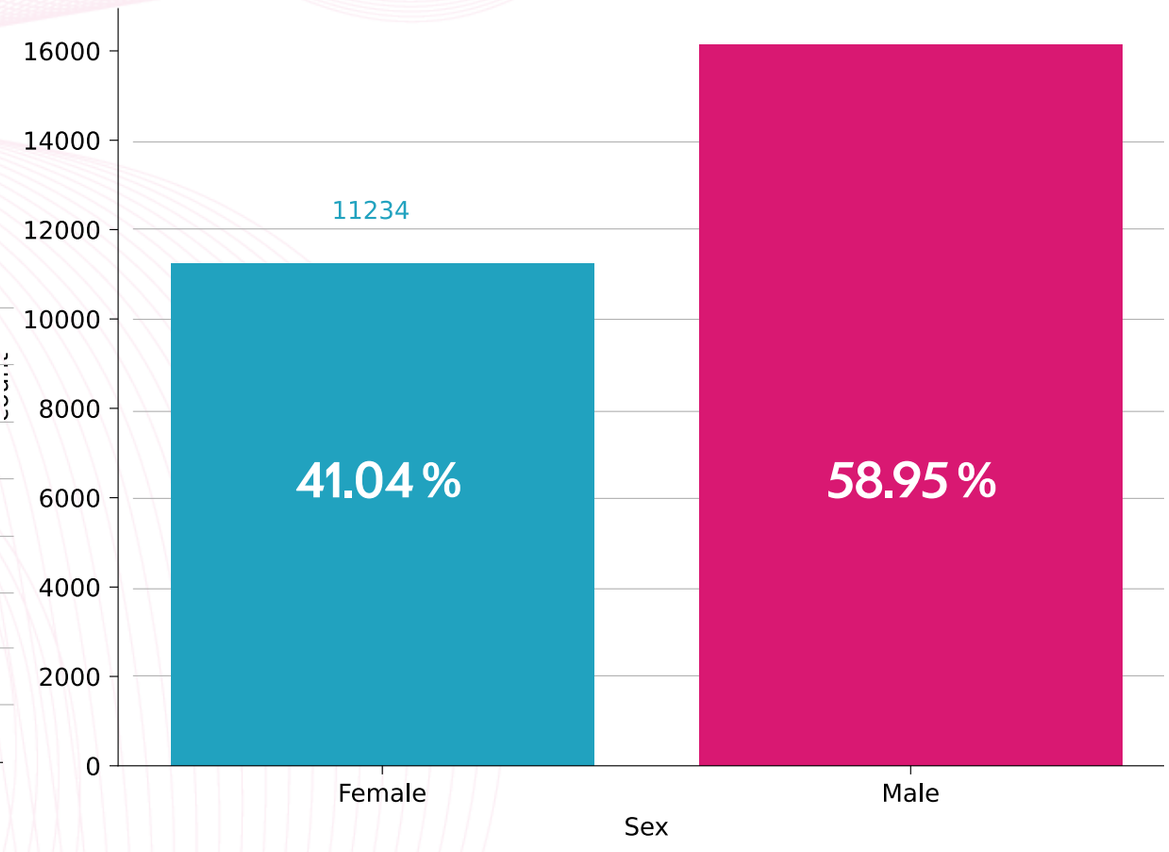
The number of people in the data that has gotten heart disease

Original Dataset





Heart Disease Dataset



SIDENOTE

It is also worth to note that although the number of women are more in the original data, the number of men suffering specifically from Heart Diseases are significantly more thus men are likely to get heart disease than women

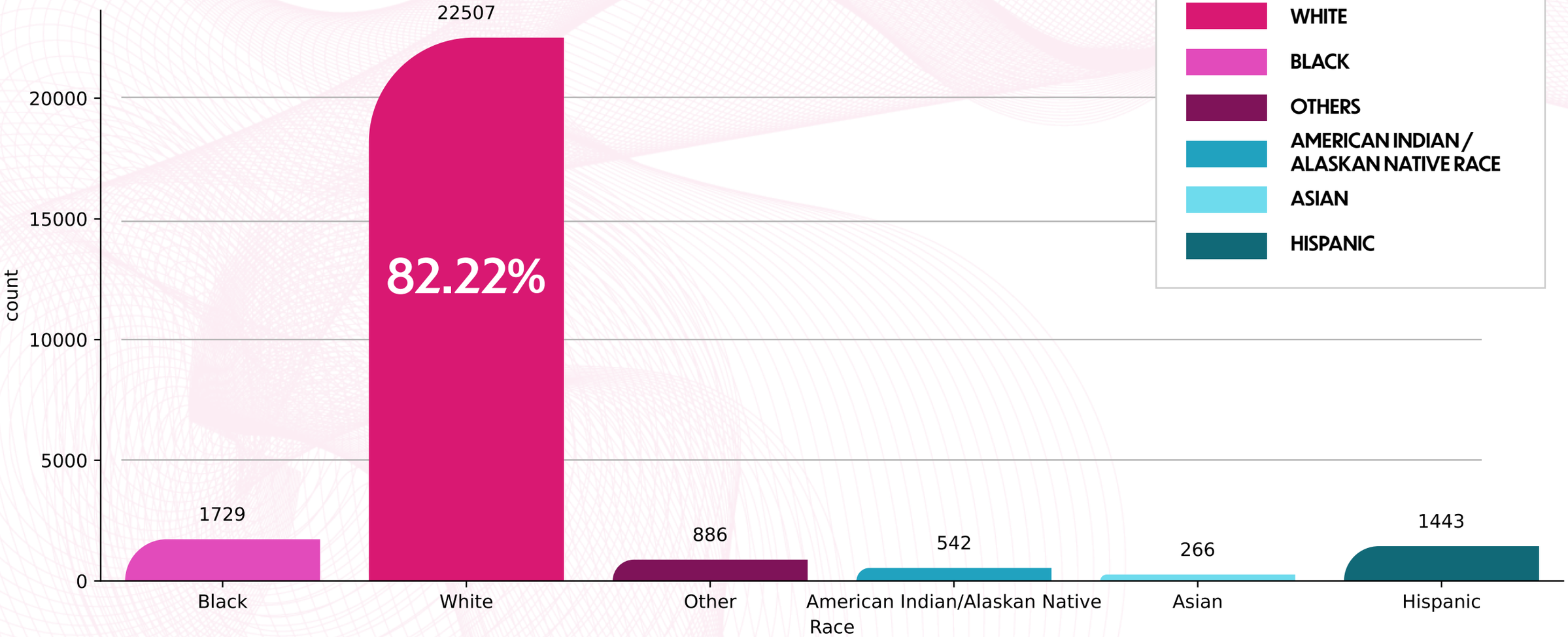
QUESTION 4

HEART DISEASE DATA

WHICH RACE HAS THE HIGHEST COUNT OF HAVING A HEART DISEASE

QUESTION 4

Which race has the highest count of having a heart disease



CONCLUSION

The highest race count for heart diseases are the white people. This value is especially high due the location of the hospital that is centrally built in a white majority neighbourhood

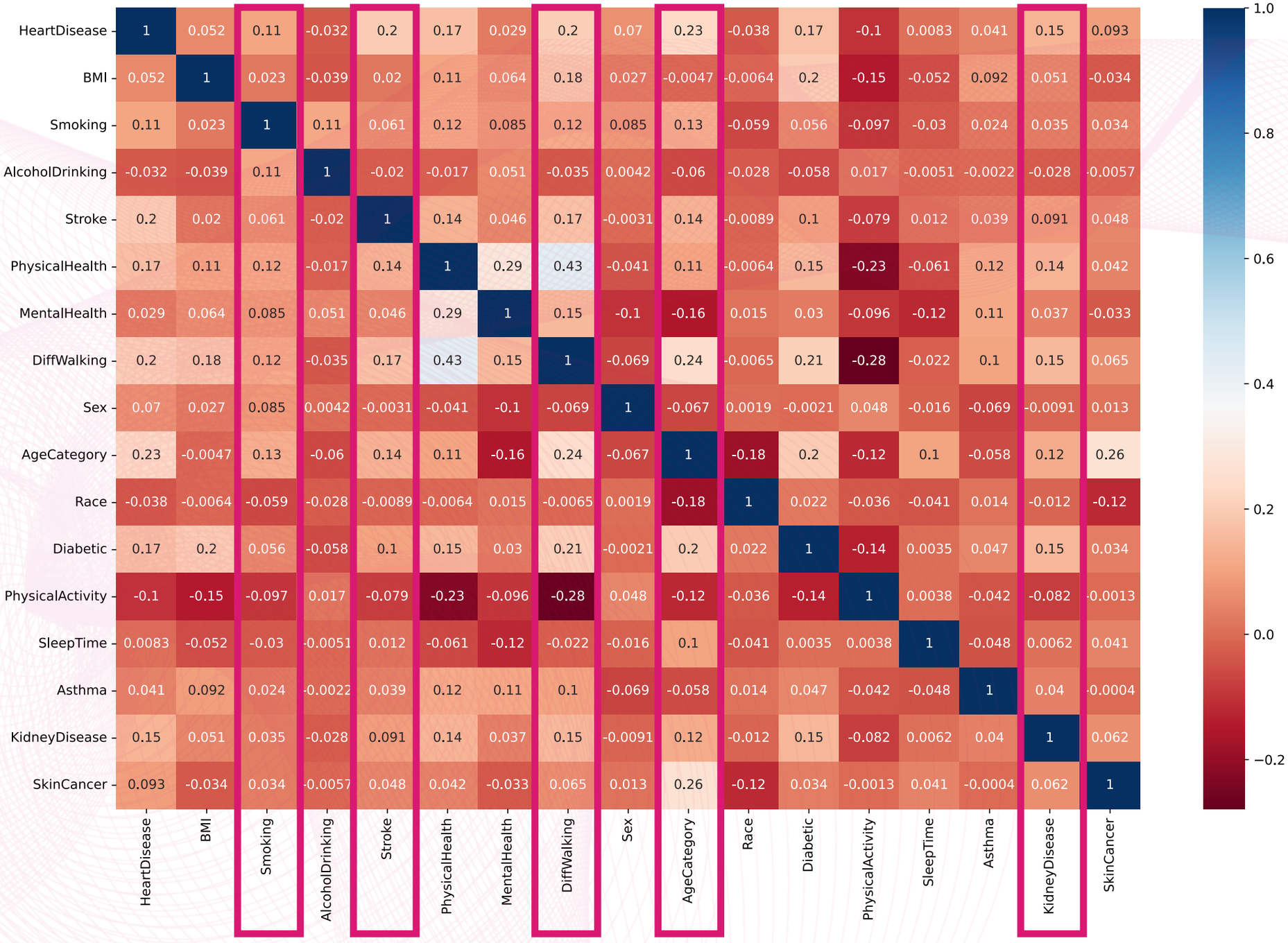
QUESTION 5

HEART DISEASE DATA

IS THERE A CORRELATION BETWEEN HEART DISEASE WITH LIFESTYLE CHOICES AND ANY PRE-EXISTING ILLNESS?

QUESTION 5

Is there a correlation between Heart Disease, lifestyle choices and any pre-existing illness?



CONCLUSION

The data seems to suggest there isn't any major correlation between the pre-existing illness or bad lifestyle choices. Although minute, bad habits of smoking, preexisting stroke & kidney disease patients and elderly patients with difficulty to walk seems to be of the highest value

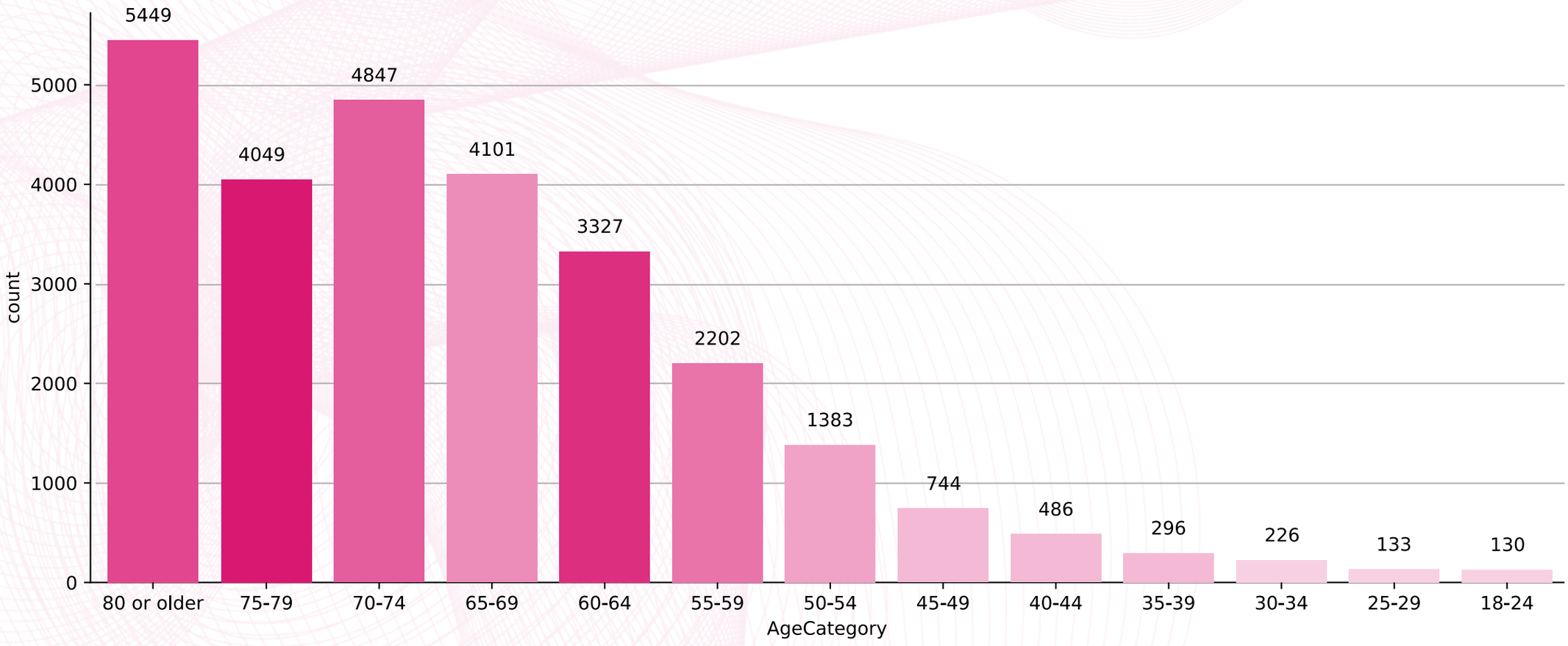
QUESTION 5

HEART DISEASE DATA

HIGHEST AGE CATEGORY THAT HAS SUFFERED FROM HEART DISEASE BEFORE

QUESTION 5

Highest age category that has suffered from heart disease before?



CONCLUSION

Data suggest that from the age of 40 onwards, the chances of a person having a Heart Disease doubles

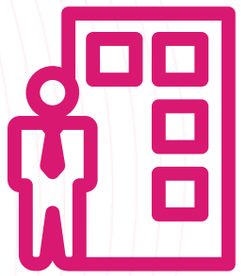
SUMMARY OF EDA

CONCLUSION

More Male suffer from Heart Disease @ 58.95%

Data Set is unbalance

The highest race count we had over all the heart disease cases where the whites at a 82% of the dataset



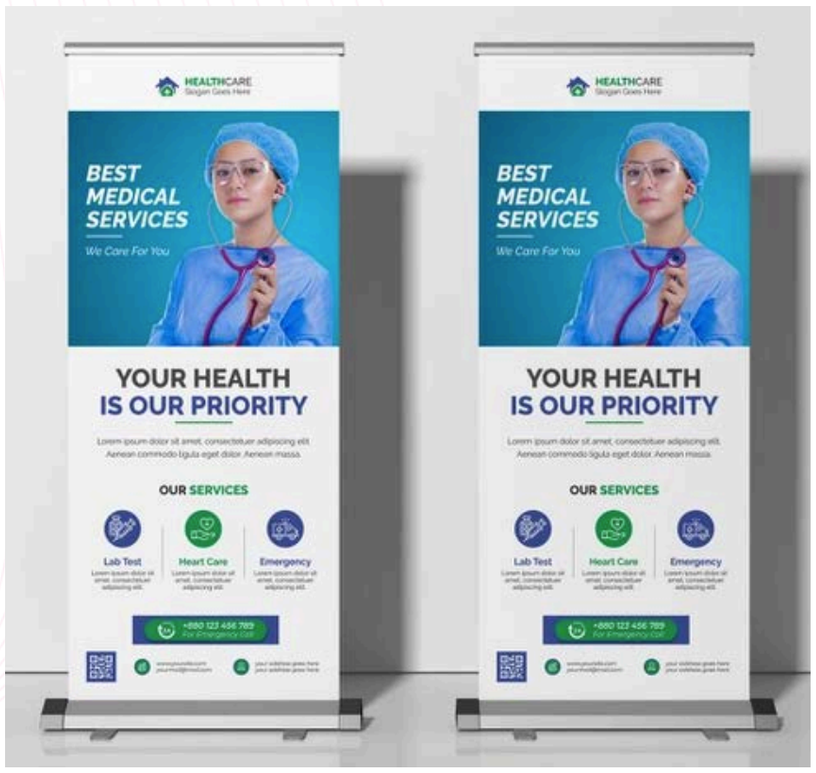
Age of 40 onwards, the chances of somebody with a similar lifestyle habits grows by 2x

OUR SUGGESTIONS BASED ON FINDINGS

Create a Mobile Medical Centre and place them strategical during Male dominant event such as Superbowl or major baseball league games

Create a temporary centre near to the Business District Area to access to the working adults

Information counters in the shopping mall or farmer's market



Pull up banner in malls entrances, bar, golf courses

MACHINE LEARNING

Since we are trying to predict that the outcome of a person having heart attack in a yes and no manner. We will be using Classification method .

Also we will be using the Recall metric to use for this project

SMOTE

Shape of X before SMOTE: (319795, 3)
Shape of X after SMOTE: (584844, 3)

Balance of positive and negative classes (%):
0 50.0
1 50.0
Name: HeartDisease, dtype: float64

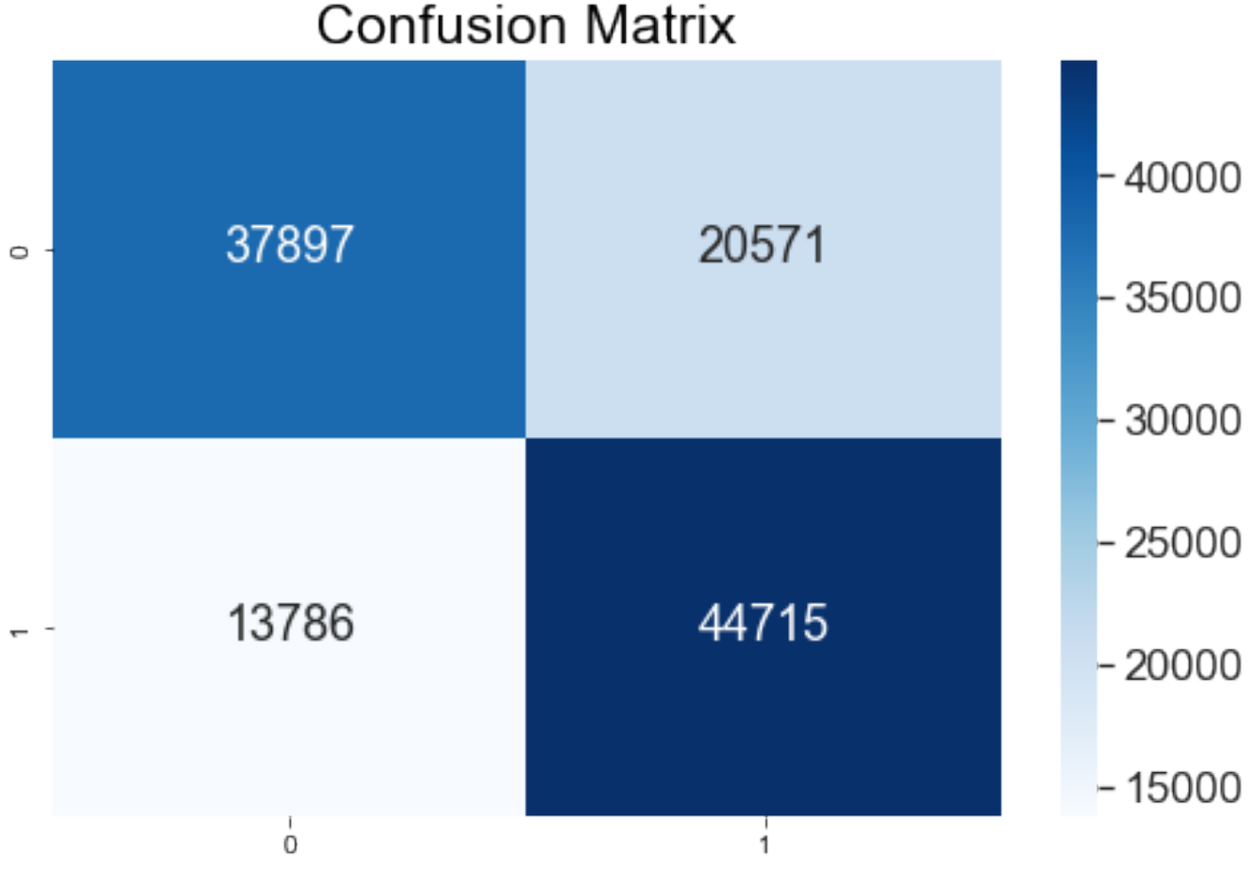
INITIAL FEATURE SELECTED

['HeartDisease', 'AgeCategory', 'Sex', 'Smoking', 'Race', 'BMI']

FINDING THE BASE MODEL

CLASSIFICATION: LOGISTIC REGRESSION

	precision	recall	f1-score	support
0	0.733259	0.648167	0.688092	58468.000000
1	0.684909	0.769446	0.722451	58501.000000
accuracy	0.706273	0.706273	0.706273	0.706273
macro avg	0.709084	0.706256	0.705271	116969.000000
weighted avg	0.709077	0.706273	0.705276	116969.000000



LOGISTIC REGRESSION

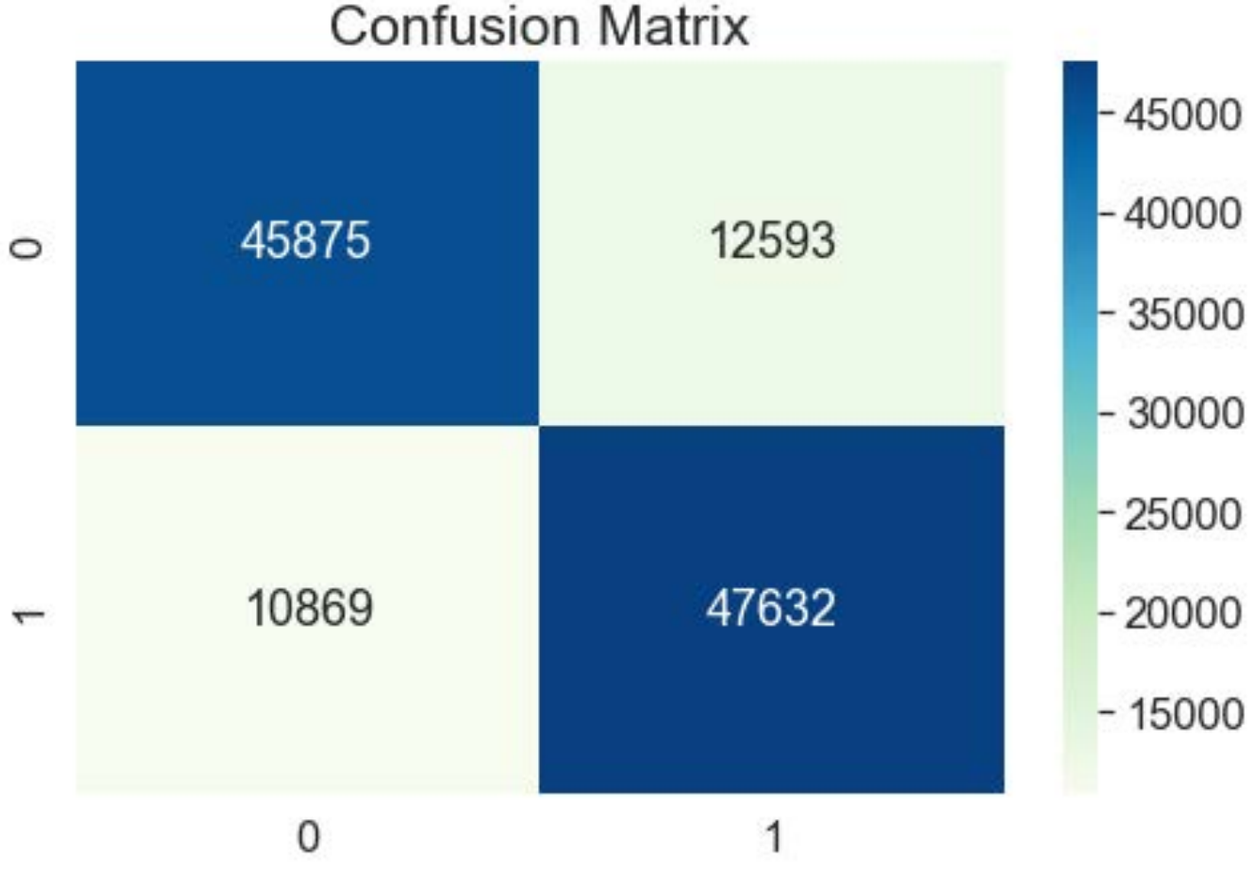
76.94%

BASE SCORE TO BEAT

FINDING THE BASE MODEL

CLASSIFICATION: DECISION TREE

	precision	recall	f1-score	support
0	0.781308	0.662003	0.716725	58468.000000
1	0.706921	0.80874807	0.757040	58501.000000
accuracy	0.738426	0.738426	0.738426	0.738426
macro avg	0.744115	0.738405	0.736882	116969.000000
weighted avg	0.744104	0.738426	0.736888	116969.000000



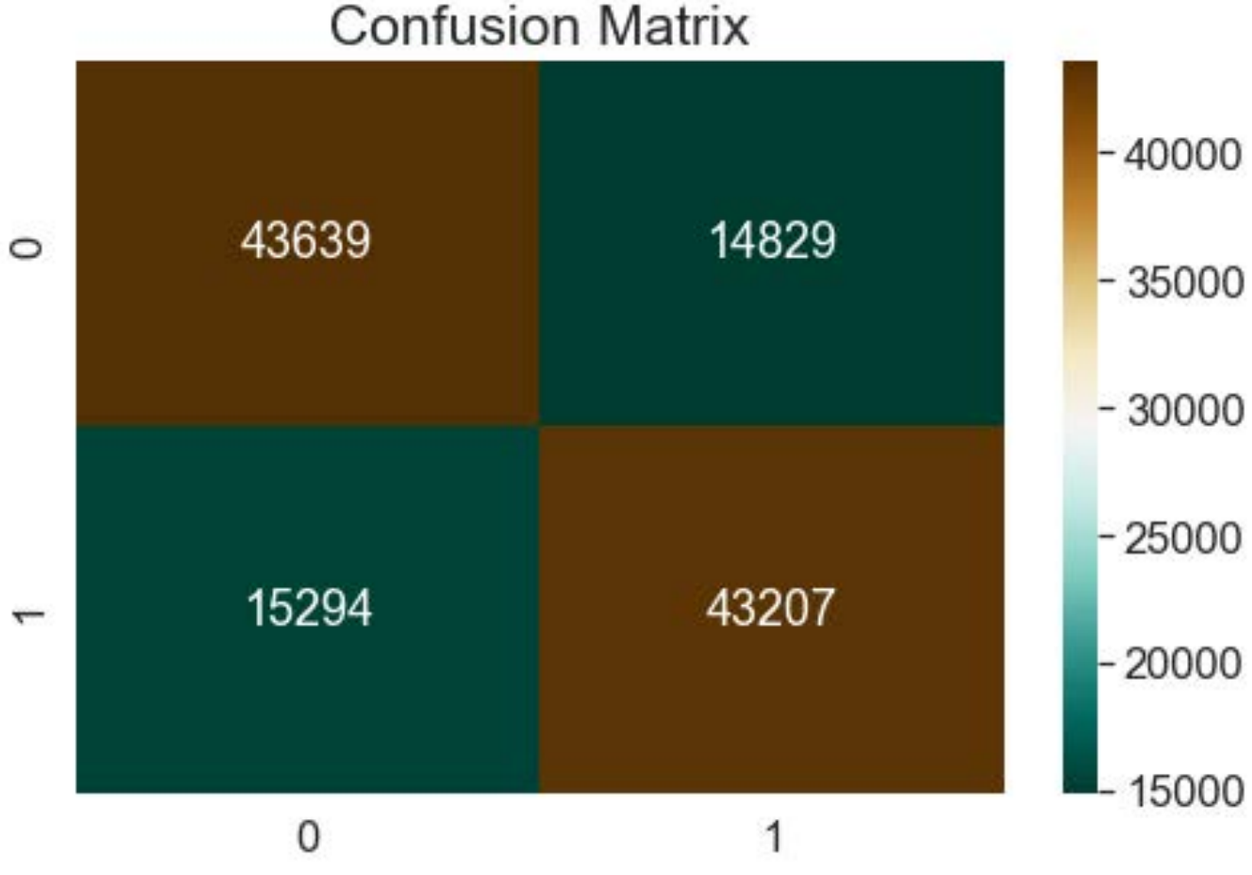
DECISION TREE

80.87%

FINDING THE BASE MODEL

CLASSIFICATION: KNN NEIGHBOUR

	precision	recall	f1-score	support
0	0.740485	0.746374	0.743418	58468.00000
1	0.744486	0.768769	0.741516	58501.00000
accuracy	0.742470	0.742470	0.742470	0.74247
macro avg	0.742486	0.742471	0.742467	116969.00000
weighted avg	0.742486	0.742470	0.742466	116969.00000



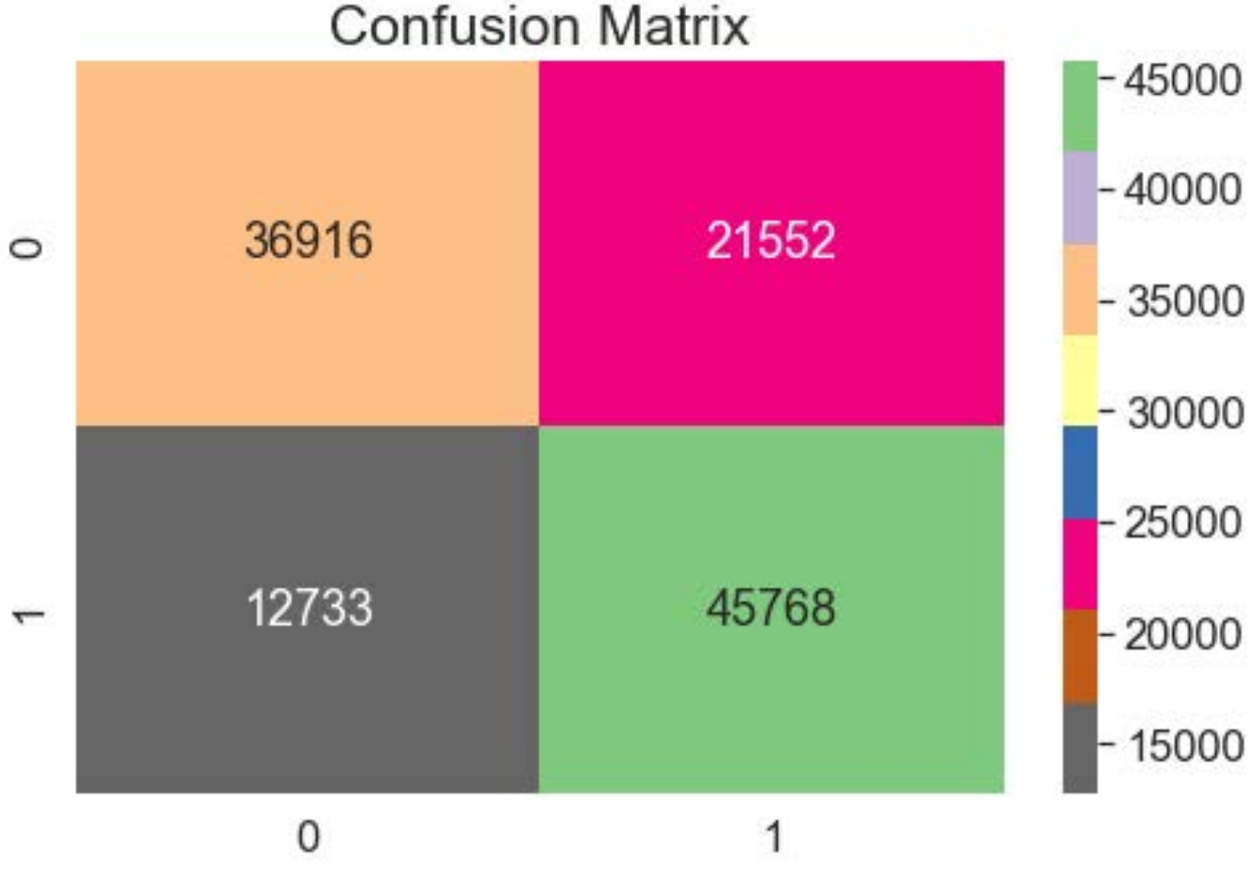
KNN NEIGHBOUR

76.87%

FINDING THE BASE MODEL

CLASSIFICATION: LINEARSVC (SVM)

	precision	recall	f1-score	support
0	0.743540	0.631388	0.682890	58468.000000
1	0.679857	0.784946	0.727510	58501.000000
accuracy	0.706888	0.706888	0.706888	0.706888
macro avg	0.711699	0.706867	0.705200	116969.000000
weighted avg	0.711690	0.706888	0.705206	116969.000000



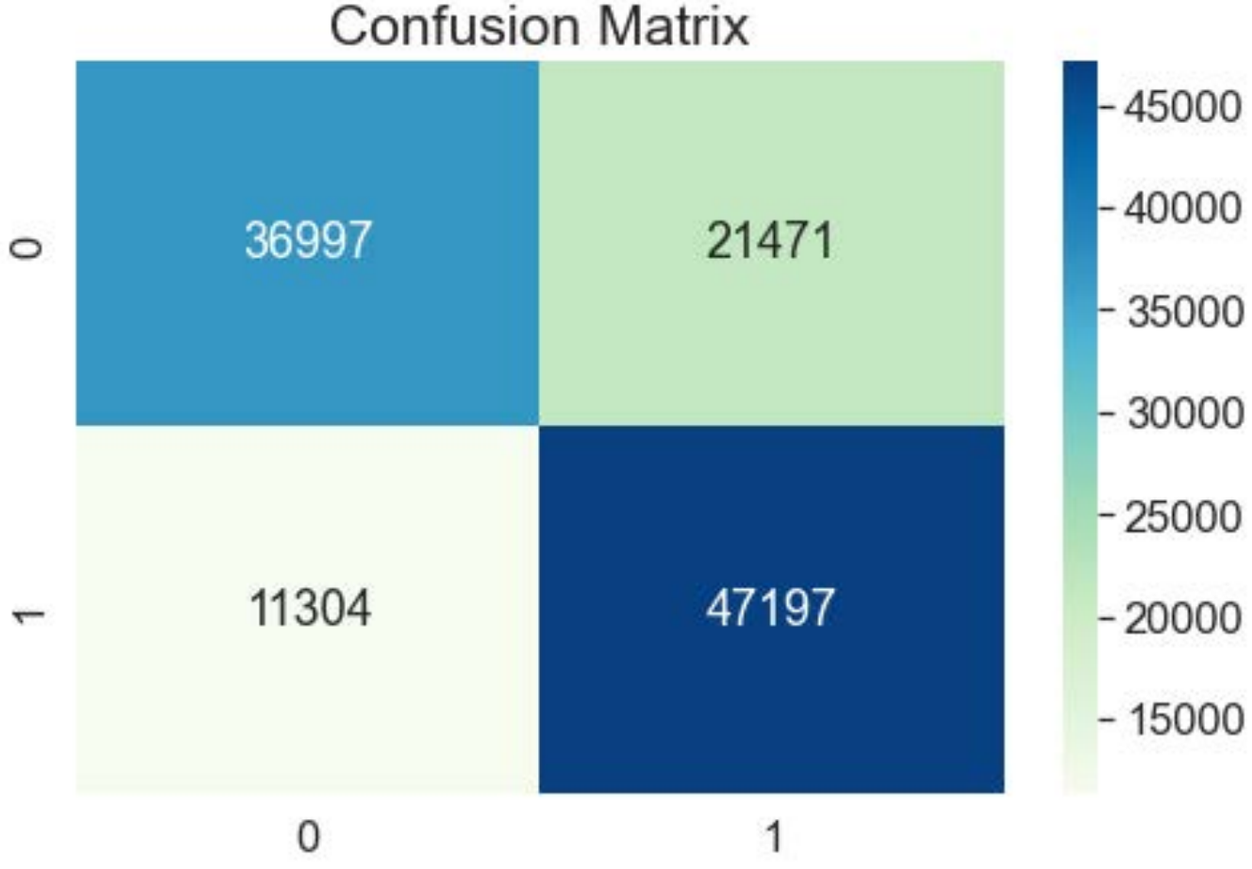
LINEARSVC

78.49%

FINDING THE BASE MODEL

CLASSIFICATION: RANDOM FOREST

	precision	recall	f1-score	support
0	0.765968	0.632773	0.693029	58468.000000
1	0.687322	0.795673	0.742272	58501.000000
accuracy	0.719798	0.719798	0.719798	0.719798
macro avg	0.726645	0.719773	0.717650	116969.000000
weighted avg	0.726633	0.719798	0.717657	116969.000000



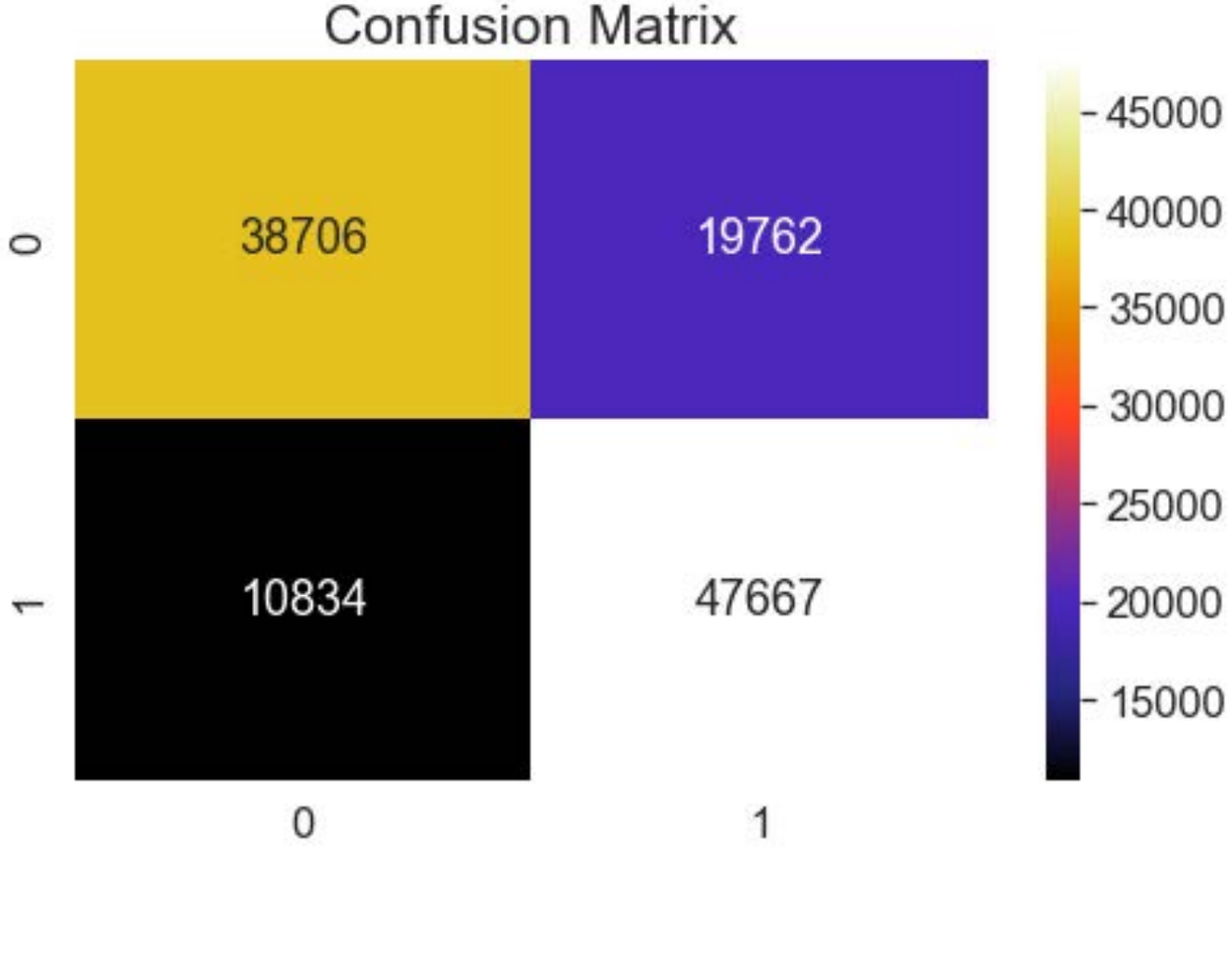
RANDOM FOREST

79.56%

FINDING THE BASE MODEL

CLASSIFICATION: XGBOOST

	precision	recall	f1-score	support
0	0.808456	0.784617	0.796358	58468.000000
1	0.790901	0.818408	0.802385	58501.000000
accuracy	0.799417	0.799417	0.799417	0.799417
macro avg	0.799678	0.799413	0.799372	116969.000000
weighted avg	0.799676	0.799417	0.799373	116969.000000



XGBOOST

81.84%

FINDING THE BASE MODEL

COMPARING CLASSIFIER SCORES

LOGISTIC REGRESSION

BASE SCORE TO BEAT 76.94%

LINEARSVC

78.49%

DECISION TREE

80.87%

RANDOM FOREST

79.56%

KNN NEIGHBOUR

76.87%

XGBOOST

HIGHEST SCORE 81.84%

As we can see the highest score goes to XGBOOST with 81.84%

MACHINE LEARNING

FEATURE SELECTION

UNIVARIATE SELECTION FOR FEATURE SELECTION

We'll then use this method to determine the features that's important, drop the features and rerun the XGBOOST Model again

	Specs	Score	
0	AgeCategory	29467.971972	SELECTED FEATURE
1	Sex	823.195671	SELECTED FEATURE
2	Smoking	2181.953023	SELECTED FEATURE
3	Race	401.672027	DROPPED
4	BMI	1224.020726	DROPPED

MACHINE LEARNING

COMPARING CLASSIFIER SCORES AFTER FEATURE SELECTION

DECREASE

LOGISTIC REGRESSION

76.94%

AFTER FT SELECTION
76.43%

KNN NEIGHBOUR

76.87%

AFTER FT SELECTION
73.86%

LINEARSVC

78.49%

AFTER FT SELECTION
78.23%

XGBOOST

81.84%

AFTER FT SELECTION
81.48%

INCREASE

RANDOM FOREST

79.56%

AFTER FT SELECTION
80.77%

DECISION TREE

80.87%

AFTER FT SELECTION
81.42%

MACHINE LEARNING

HYPER PARAMETER WITH GRID SEARCH CV

```
param_grid = {
    'max_depth': [5, 6, 7],
    'learning_rate': [0.1, 0.5, 1],
    'reg_lambda': [10.0, 20, 100],
     'scale_pos_weight': [x]
}
```

```
optimal_params = GridSearchCV(
                    estimator = xgb_grid,
                    param_grid = param_grid,
                    scoring = 'recall',
                    verbose = 0,
                    cv = 3
```

We'll then apply the parameters back into XGBOOST and rerun to get the result

```
XGBC = xgb.XGBClassifier(learning_rate=1, max_depth=7, reg_lambda=10.0, scale_pos_weight=1.0)
```

XGBOOST		
81.84%	AFTER FT SELECTION 81.48%	AFTER HYPERTUNING 82.91%

SUMMARY

THANK YOU

https://github.com/AffendiAbdullah/machine_learning_heart_disease_final_project.git