**Objective**

**The objective of our Project is building a prediction model for heart attack using the heart attack and prediction dataset provided by Kaggle.**

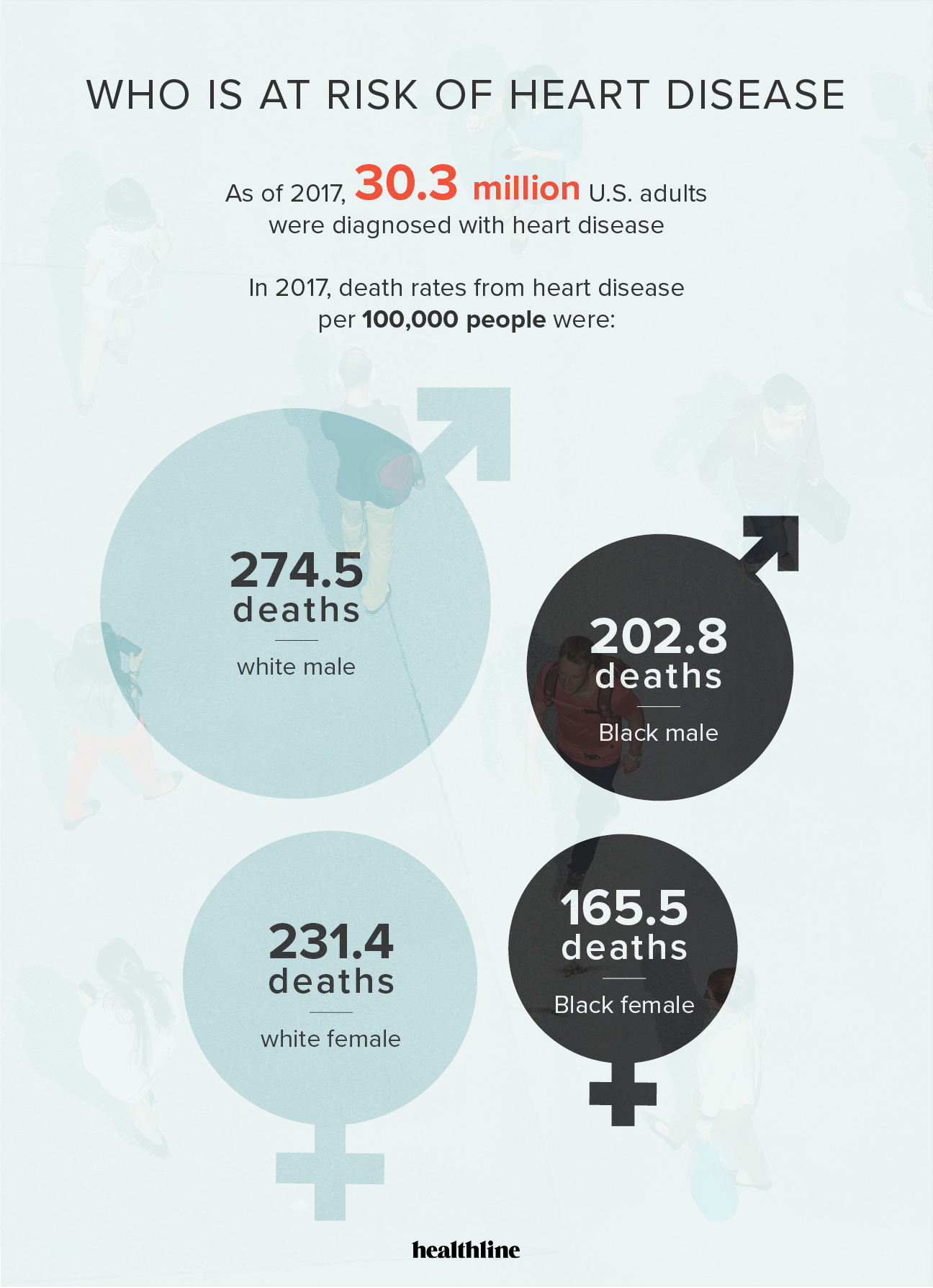
**Statistics**

According to the [Centers for Disease Control and Prevention (CDC)](https://www.cdc.gov/heartdisease/facts.htm), approximately every 40 seconds an American will have a heart attack. Every year, 805,000 Americans have a heart attack, 605,000 of them for the first time.

About [12 percent](https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2728009?resultClick=1)  of people who have a heart attack will die from it.

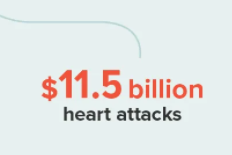
According to the [American Heart Association](https://www.ahajournals.org/doi/pdf/10.1161/cir.0000000000000351?sid=beb5f268-4205-4e62-be8f-3caec4c4d9b7&), 26 percent of women will die within a year of a heart attack compared with 19 percent of men.

By 5 years after a heart attack, almost 50 percent of women die, develop heart failure, or have a stroke compared with 36 percent of men.



**How much does it cost?**

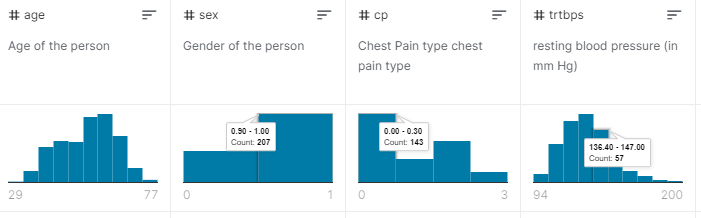
According to the CDC, the number of emergency room visits in 2017 for issues related to the heart and blood vessels was nearly [5 million](https://www.cdc.gov/nchs/data/nhamcs/web_tables/2017_ed_web_tables-508.pdf) . In 2016, [72 million](https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2016_namcs_web_tables.pdf)  people made heart disease-related visits to their doctors.

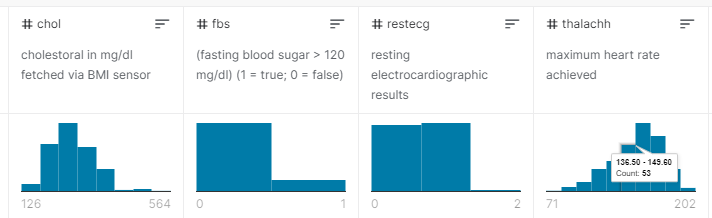


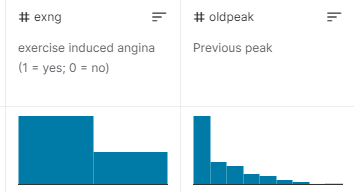
**Why this model**

After looking at all these data we can see, there is a larger need of a prediction model which can predict heart attack in individuals based on the data which includes age, sex, chest pain type, resting blood pressure etc. So, for building our model we have used data set from Kaggle which includes data for 303 individuals using which we can predict the heart attack rate.

**Dataset Details**







For building the prediction model we have chosen AWS because it offers the broadest and deepest set of [machine learning services](https://aws.amazon.com/machine-learning/#Explore_AWS_Machine_Learning_services) like  SageMaker notebook instance, S3 bucket for deploying dataset and supporting cloud [infrastructure](https://aws.amazon.com/machine-learning/infrastructure/?c=ml&sec=int), putting machine learning in the hands of every developer, data scientist and expert practitioner.

**The tools used for this project mainly consisted of AWS services like Amazon SageMaker Studio and S3 Bucket.**

Initially the project seemed easy to implement as the data given pretty accurate and it required less cleaning also the expectations were straightforward. Although this was easy to implement in a local environment, implementing it using Cloud services was a task. Going with the suggestion, we decided to use Amazon SageMaker Studio and started our research. Since Amazon SageMaker is a very vast service covering many aspects of Machine Learning, it was difficult to grasp the concept. After researching for a few days we learned that Amazon SageMaker is a service by AWS for building, training and deploying machine learning models. SageMaker Studio would let us decide the type of machine we prefered so that there was no need to manage any complex AMIs or security groups. This made it very easy to get started. SageMaker also provided access to GPUs and big machines with high amounts of RAM that might not be possible on a local setup like EC2 hence we decided to go ahead with SageMaker.

Another advantage we found while using it was that it comes with environments pre-configured, that is we did not need to install XGBoost, SciKit Learn or other common libraries separately.

Since data was already provided to us, initially we directly uploaded it in Jupyter Notebooks and started working with it. Later after reading the sensitivity about medical data and it’s protection by HIPAA (The Health Insurance Portability and Accountability Act of 1996 is a United States federal statute enacted by the 104th United States Congress and signed into law by President Bill Clinton on August 21, 1996.) We decided to use a different and more secure storage. That’s when we decided to go with Amazon Simple Storage Service. S3  is an object storage service that offers industry-leading scalability, data availability, security, and performance. This would ensure that the data is protected from public access and is readily accessible.

We created a Jupyter Notebook instance from SageMaker and selected Any instance of S3 bucket. The notebook access was given to the root user in the IAM permissions and encryption and deployed the instance. The Conda environment was selected in Jupyter Notebooks and imported all the necessary libraries like pandas, numpy, matplotlib, Scikit Learn etc. Since we were using SageMaker all these were preinstalled.

Then we accessed the S3 bucket through python code to store the data. This data was then downloaded to the SageMaker instance and loaded in a dataframe to be divided into training and test sets (70 to 30 ratio).

After creating and training the model we stored it in the S3 bucket. This model was Naive Bayes. Inorder get the predictions this trained model would be fed data to deliver the prediction.

Reference –

<https://www.healthline.com/health/heart-disease/statistics#Who-is-at-risk>?

<https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>