HEART DISEASE PREDICTION

HEALTHCARE DATA ANALYTICS

HINF 6400

FALL 2023

PRESENTED TO:PROFESSOR DAN RIES

EXECUTIVE SUMMARY

- Team's mission statement
- Background / Identifying the opportunity statement
- Statistical relevance of heart disease in the U.S
- Plan of Analysis
- Data Analytics preparation approach (EDA, statistical significance, Key variable identification, model implementation)
- Dataset Description
- Relative impact of different independent variables
- Key predictive variables / risk factors identified
- Data visualizations
- *Running the Analytics* Statistical and ML Predictive modeling implementation
- *RESULTS* Findings , conclusion and recommendations

TEAM MISSION STATEMENT

"Maximizing human longevity, healthier communities and optimal health outcomes through data-driven patient-centered Healthcare Analytics"

The Team

\

OPPORTUNITY STATEMENT

- *"Through the implementation of Health Analytics tools and with a healthcare data driven mindset for bettering patient health outcomes, the team intends to implement several ML tools that will allow for a better understanding of what are the independent variables and risk factors, that are more relevant in predicting heart disease.*
- *The target is a binary classification where "0" indicates no heart disease and "1" indicates the presence of heart disease. With all healthcare stakeholders in mind; Doctors, Clinical and medical personal, impacted communities and the patients themselves, the team will present its findings regarding heart disease and will provide a set of data driven, research and evidence recommendations.*

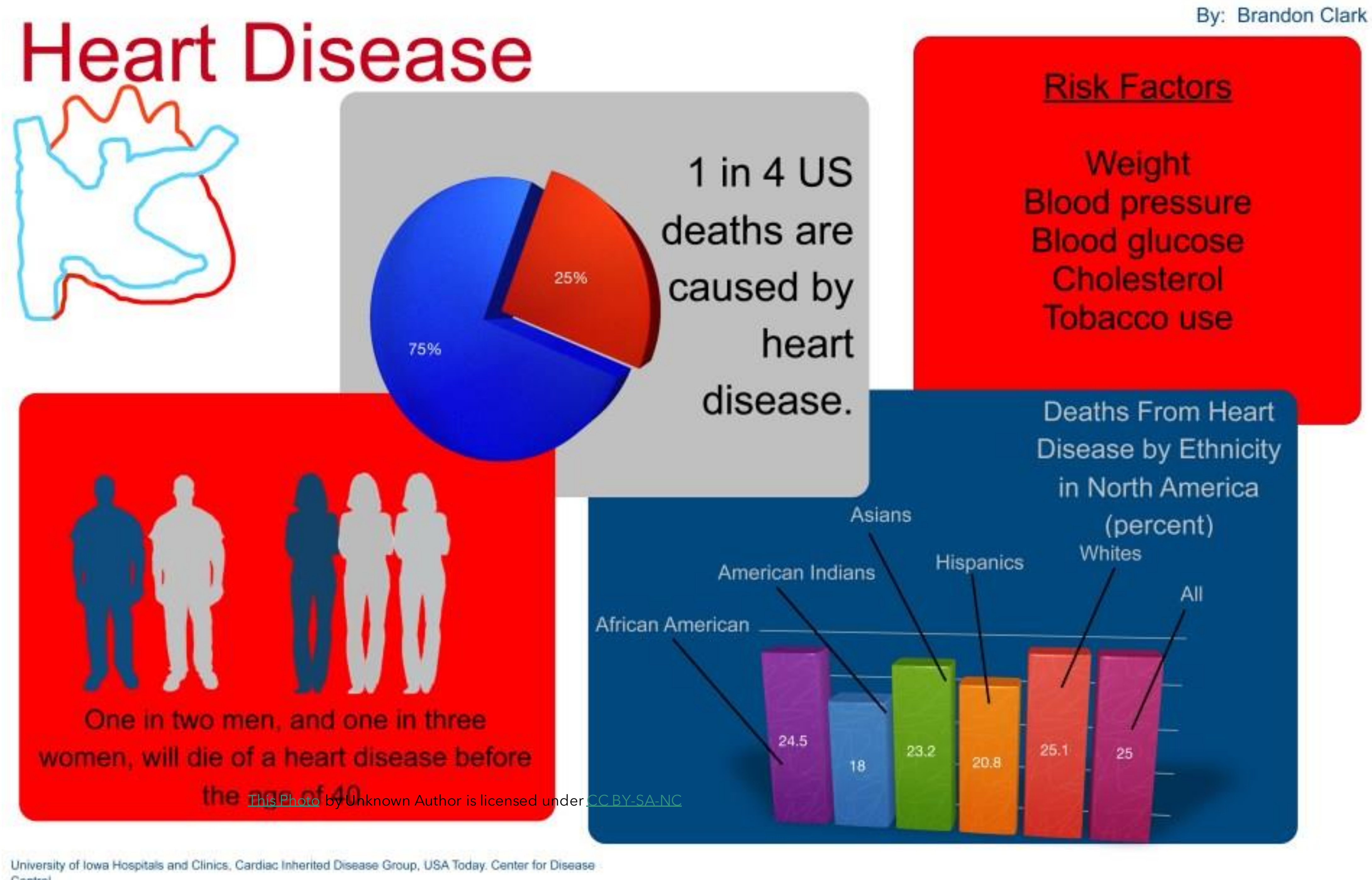
RELEVANT STATISTICS OF HEART DISEASE IN THE U.S

- Heart disease is the **leading cause of death** for men, women and people of most racial and ethnic groups in the United States.
- One person dies **every 33 seconds** from heart disease in the U.S
- **699,659 people** in the U.S died from **heart disease** in the U.S in 2022
- 607,790 people in the U.S died from Cancer in the U.S in 2022
- **1 of every 5 deaths** in the U.S is related to heart disease

Heart disease costs the United states approx. **$ 300 BILLION a year !**

(Healthcare services, medicines, and lost productivity due to death)

Source: https://www.cdc.gov Provisional Mortality Data--- United Stares 2022



By: Brandon Clark

Heart Disease

1 in 4 US deaths are caused by heart disease.

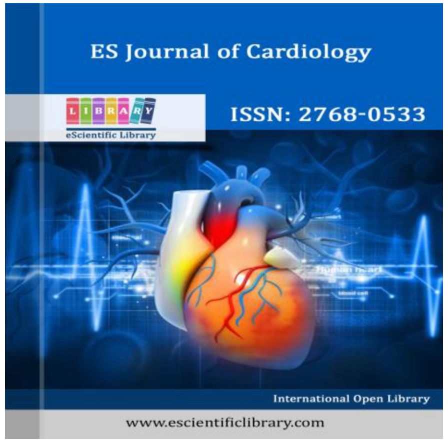
Risk Factors

Weight
Blood pressure
Blood glucose
Cholesterol
Tobacco use

One in two men, and one in three women, will die of a heart disease before the age of 40

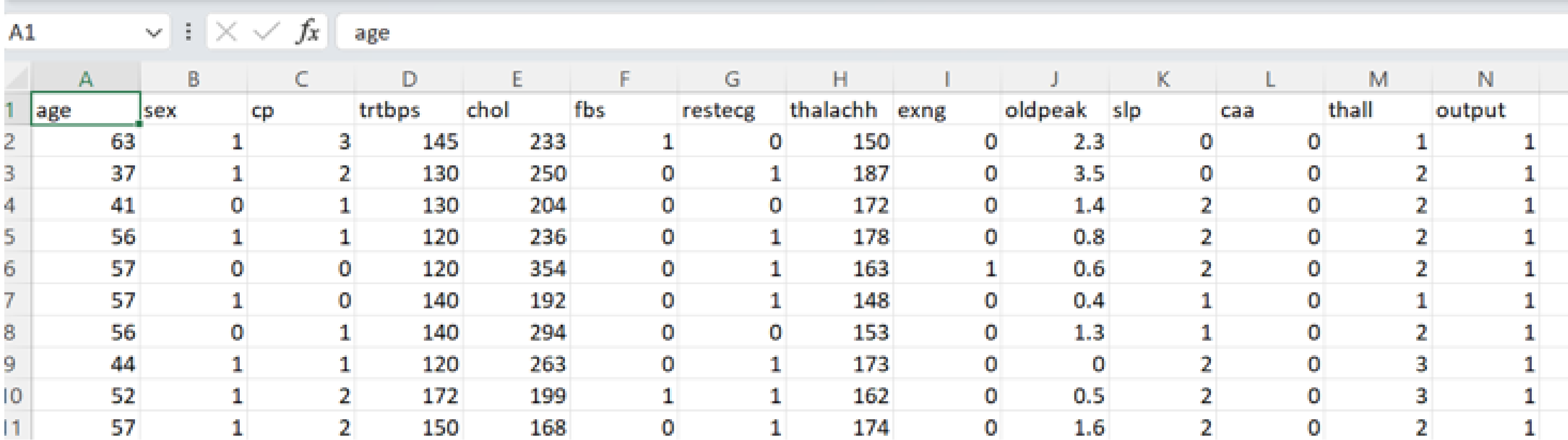
Deaths From Heart Disease by Ethnicity in North America (percent)

University of Iowa Hospitals and Clinics, Cardiac Inherited Disease Group, USA Today, Center for Disease Control



DATASET DESCRIPTION

The data set/s we intend to use. The data set/s we will be using were obtained from Kaggle (data_heart). The name of the data set is ***Heart Attack & Prediction Dataset*** available at www.kaggle.com



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

This Photo by Unknown Author is li

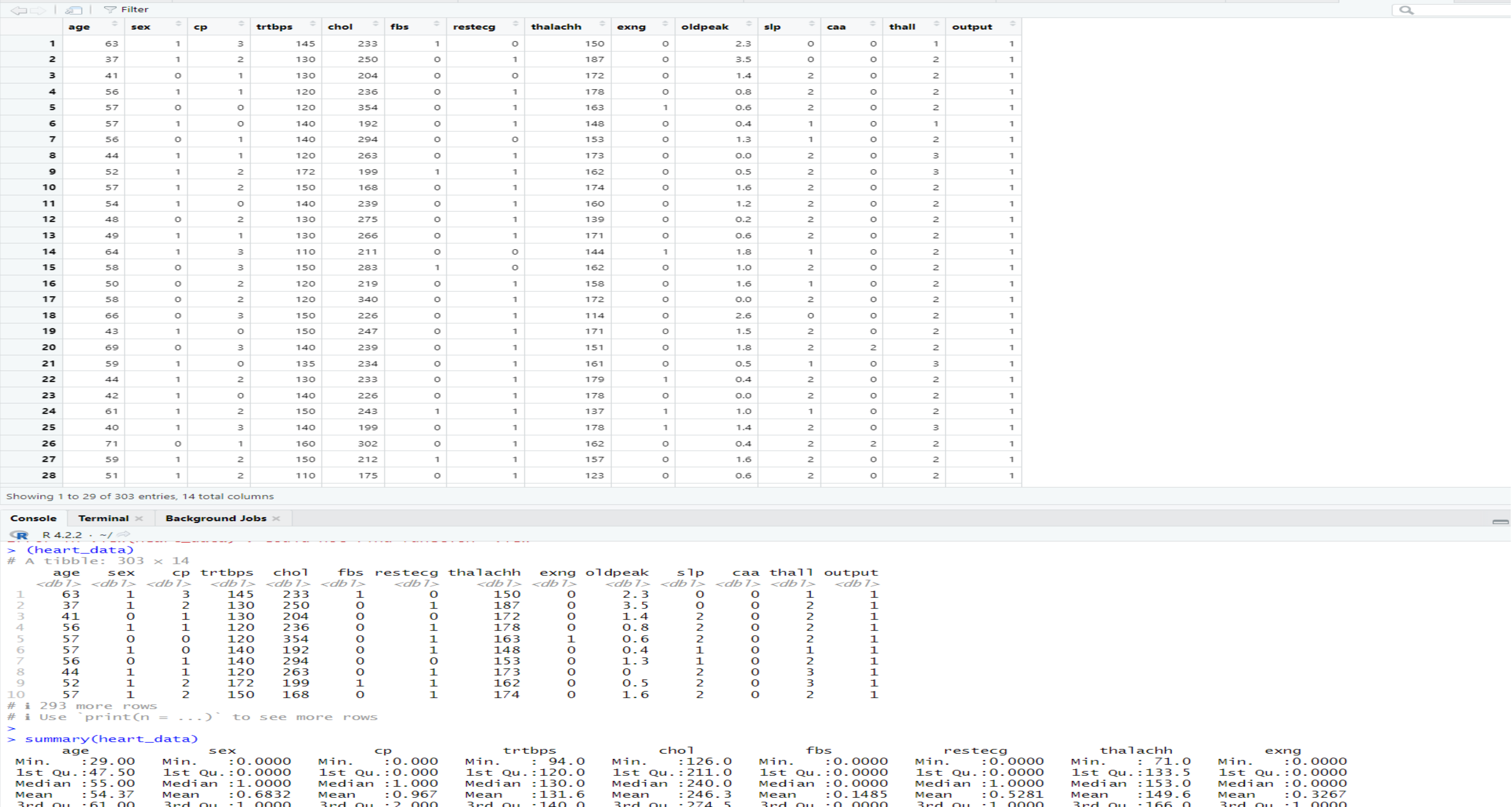
VARIABLE DESCRIPTION / H.D RELEVANCE

- Age: The age of the individual collected during data collection.
- Sex: Sex of the patient
- Chest Pain Type (cp): Describes the type of chest pain experienced by the patients
- Resting Blood Pressure (trtbps)
- Cholesterol (chol)
- Fasting Blood Sugar (fbs)
- Resting Electrocardiographic Results (restecg)
- Thallium Stress Test Result (thalach)
- Exercise-Induced Angina (exng)
- ST Depression Induced by Exercise Relative to Rest (oldpeak)
- Slope of the Peak Exercise ST Segment (slp)
- Number of Major Vessels (caa):
- Output

COMPREHENSIVE DATA ANALYSIS OVERVIEW

- EDA exploratory data analysis
- Important statistics CAD (# cause of death in the U.S)
- Statistical significance (The dataset / data needs)
- Outlier identification (non-existing) All 303 observations and 13 variables complete
- Presence of missing values (None)
- Data visualizations (Summary statistics, population comparison, correlations, modeling)
- ML model implementation (research and data needs oriented)
- Findings and conclusions (Comprehensive approach to data analytics)
- The stakeholders and the patient's voice as the main drivers of health outcomes

VISUALIZING AND QUANTIFYING OUR DATA(

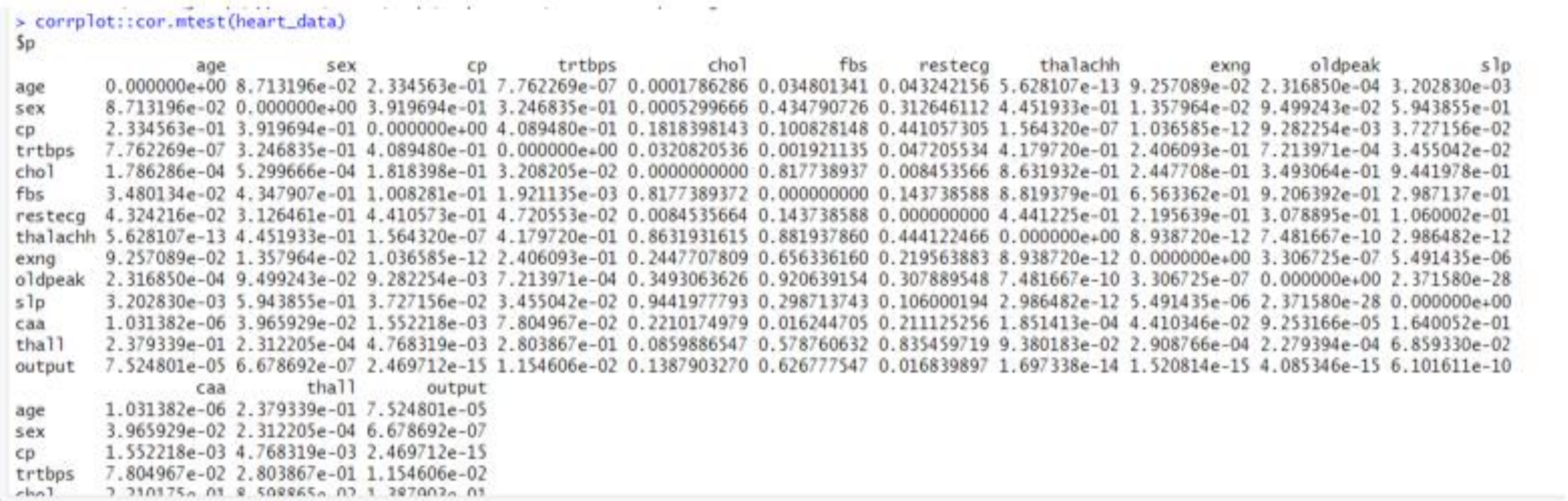


	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
9	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
10	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
11	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
12	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
13	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
14	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
15	58	0	3	150	283	1	0	162	0	1.0	2	0	2	1
16	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
17	58	0	2	120	340	0	1	172	0	0.0	2	0	2	1
18	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
19	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
20	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
21	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
22	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
23	42	1	0	140	226	0	1	178	0	0.0	2	0	2	1
24	61	1	2	150	243	1	1	137	1	1.0	1	0	2	1
25	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
26	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
27	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
28	51	1	2	110	175	0	1	123	0	0.6	2	0	2	1

Showing 1 to 29 of 303 entries, 14 total columns

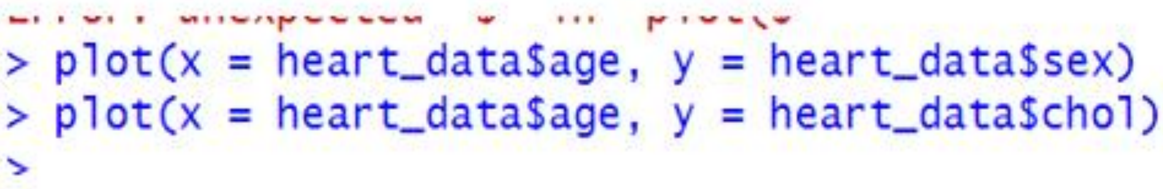
```
> (heart_data)
# A tibble: 303 × 14
     age   sex    cp trtbps  chol   fbs restecg thalachh  exng oldpeak   slp   caa thall output
   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl>
 1    63     1     3    145   233     1       0      150     0     2.3     0     0     1      1
 2    37     1     2    130   250     0       1      187     0     3.5     0     0     2      1
 3    41     0     1    130   204     0       0      172     0     1.4     2     0     2      1
 4    56     1     1    120   236     0       1      178     0     0.8     2     0     2      1
 5    57     0     0    120   354     0       1      163     1     0.6     2     0     2      1
 6    57     1     0    140   192     0       1      148     0     0.4     1     0     1      1
 7    56     0     1    140   294     0       0      153     0     1.3     1     0     2      1
 8    44     1     1    120   263     0       1      173     0     0       2     0     3      1
 9    52     1     2    172   199     1       1      162     0     0.5     2     0     3      1
10    57     1     2    150   168     0       1      174     0     1.6     2     0     2      1
# ℹ 293 more rows
# ℹ Use `print(n = ...)` to see more rows
>
> summary(heart_data)
      age             sex               cp             trtbps           chol            fbs              restecg          thalachh          exng
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0   Min.   :0.0000
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000   Median :1.0000   Median :153.0   Median :0.0000
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6   Mean   :0.3267
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000
```

CORRELATION AMONG VARIABLES



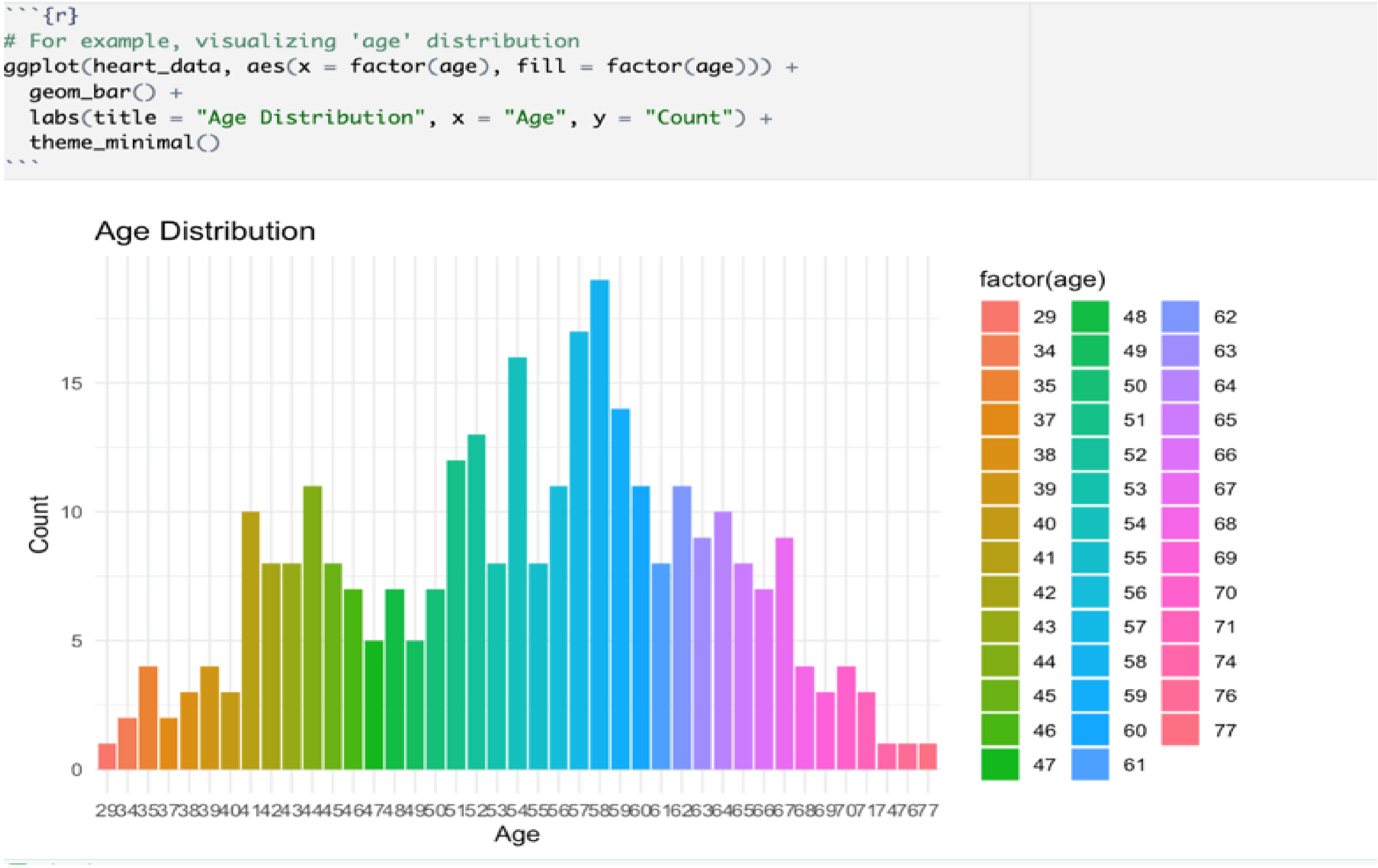
```
> corrplot::cor.mtest(heart_data)
$p
                  age          sex           cp       trtbps         chol         fbs     restecg     thalachh         exng      oldpeak          slp
age      0.000000e+00 8.713196e-02 2.334563e-01 7.762269e-07 0.0001786286 0.034801341 0.043242156 5.628107e-13 9.257089e-02 2.316850e-04 3.202830e-03
sex      8.713196e-02 0.000000e+00 3.919694e-01 3.246835e-01 0.0005299666 0.434790726 0.312646112 4.451933e-01 1.357964e-02 9.499243e-02 5.943855e-01
cp       2.334563e-01 3.919694e-01 0.000000e+00 4.089480e-01 0.1818398143 0.100828148 0.441057305 1.564320e-07 1.036585e-12 9.282254e-03 3.727156e-02
trtbps   7.762269e-07 3.246835e-01 4.089480e-01 0.000000e+00 0.0320820536 0.001921135 0.047205534 4.179720e-01 2.406093e-01 7.213971e-04 3.455042e-02
chol     1.786286e-04 5.299666e-04 1.818398e-01 3.208205e-02 0.0000000000 0.817738937 0.008453566 8.631932e-01 2.447708e-01 3.493064e-01 9.441978e-01
fbs      3.480134e-02 4.347907e-01 1.008281e-01 1.921135e-03 0.8177389372 0.000000000 0.143738588 8.819379e-01 6.563362e-01 9.206392e-01 2.987137e-01
restecg  4.324216e-02 3.126461e-01 4.410573e-01 4.720553e-02 0.0084535664 0.143738588 0.000000000 4.441225e-01 2.195639e-01 3.078895e-01 1.060002e-01
thalachh 5.628107e-13 4.451933e-01 1.564320e-07 4.179720e-01 0.8631931615 0.881937860 0.444122466 0.000000e+00 8.938720e-12 7.481667e-10 2.986482e-12
exng     9.257089e-02 1.357964e-02 1.036585e-12 2.406093e-01 0.2447707809 0.656336160 0.219563883 8.938720e-12 0.000000e+00 3.306725e-07 5.491435e-06
oldpeak  2.316850e-04 9.499243e-02 9.282254e-03 7.213971e-04 0.3493063626 0.920639154 0.307889548 7.481667e-10 3.306725e-07 0.000000e+00 2.371580e-28
slp      3.202830e-03 5.943855e-01 3.727156e-02 3.455042e-02 0.9441977793 0.298713743 0.106000194 2.986482e-12 5.491435e-06 2.371580e-28 0.000000e+00
caa      1.031382e-06 3.965929e-02 1.552218e-03 7.804967e-02 0.2210174979 0.016244705 0.211125256 1.851413e-04 4.410346e-02 9.253166e-05 1.640052e-01
thall    2.379339e-01 2.312205e-04 4.768319e-03 2.803867e-01 0.0859886547 0.578760632 0.835459719 9.380183e-02 2.908766e-04 2.279394e-04 6.859330e-02
output   7.524801e-05 6.678692e-07 2.469712e-15 1.154606e-02 0.1387903270 0.626777547 0.016839897 1.697338e-14 1.520814e-15 4.085346e-15 6.101611e-10
                  caa        thall       output
age      1.031382e-06 2.379339e-01 7.524801e-05
sex      3.965929e-02 2.312205e-04 6.678692e-07
cp       1.552218e-03 4.768319e-03 2.469712e-15
trtbps   7.804967e-02 2.803867e-01 1.154606e-02
chol     2.210175e-01 8.598865e-02 1.387903e-01
```

The function `> corrplot::cor.mtest(heart_data)` above show the correlations among different variables, allowing to see which are the stronger predictor variables of heart disease. The stronger the correlation is to 1 the higher the predictive strength of that variable.



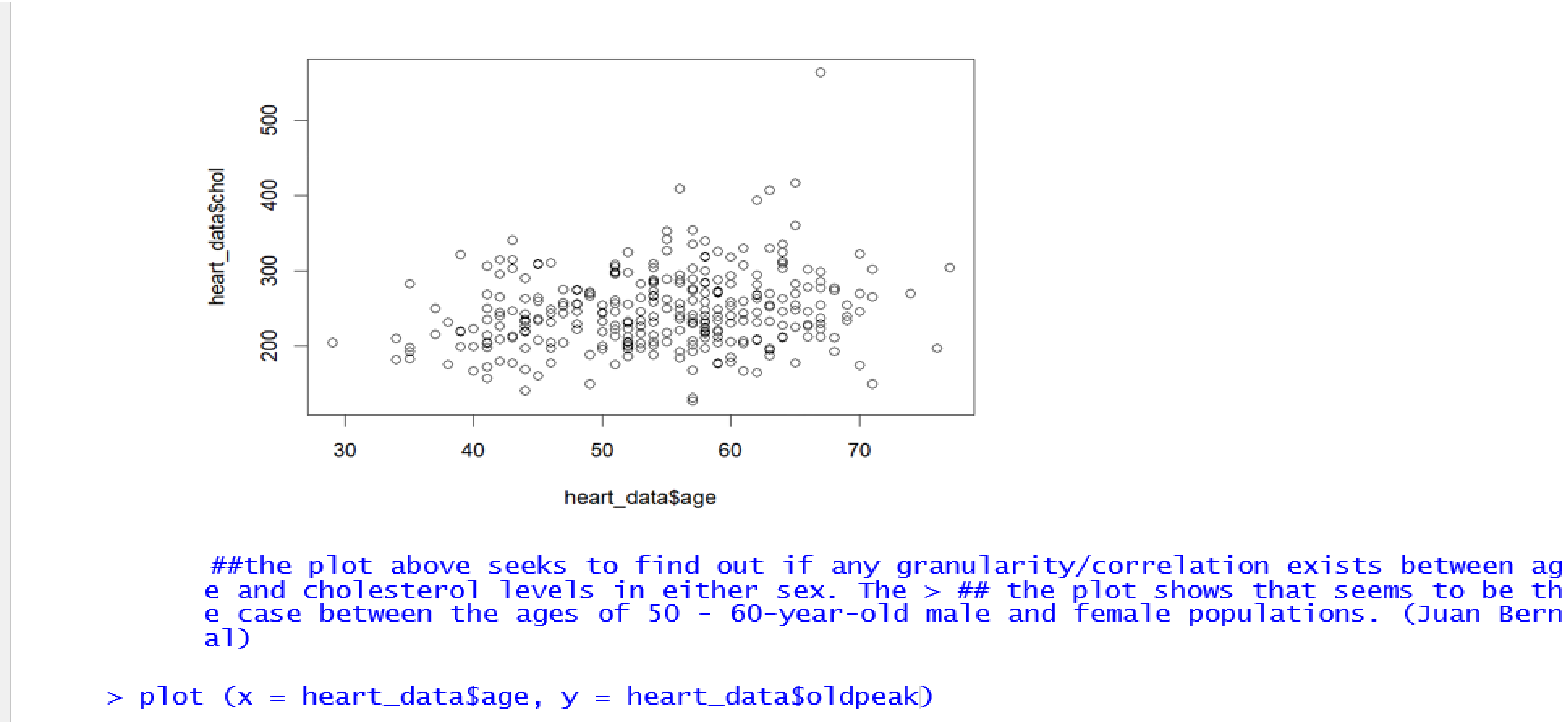
```
Error: unexpected '$' in "plot($"
> plot(x = heart_data$age, y = heart_data$sex)
> plot(x = heart_data$age, y = heart_data$chol)
>
```

AGE DISTRIBUTION



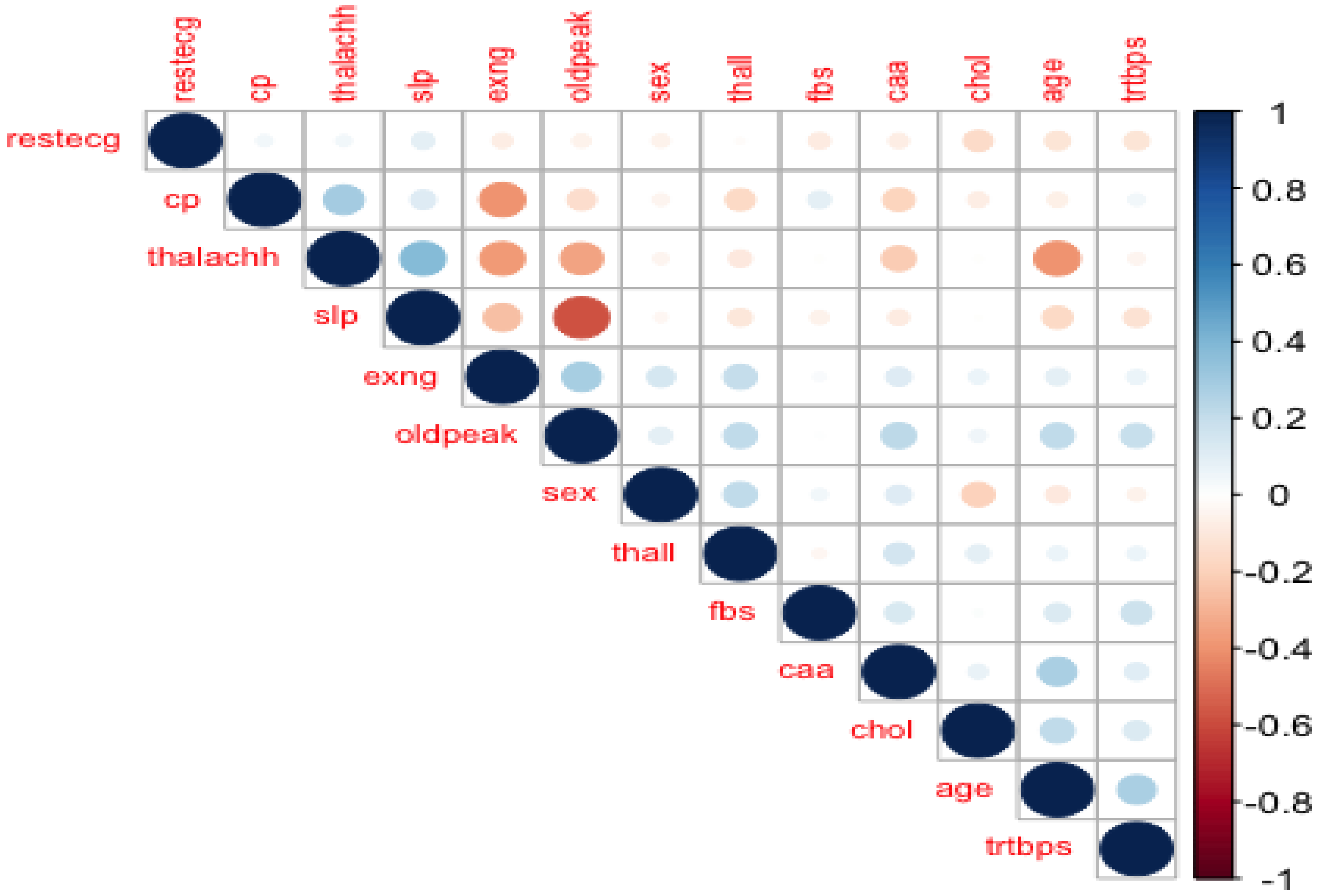
```{r}
# For example, visualizing 'age' distribution
ggplot(heart_data, aes(x = factor(age), fill = factor(age))) +
  geom_bar() +
  labs(title = "Age Distribution", x = "Age", y = "Count") +
  theme_minimal()
```

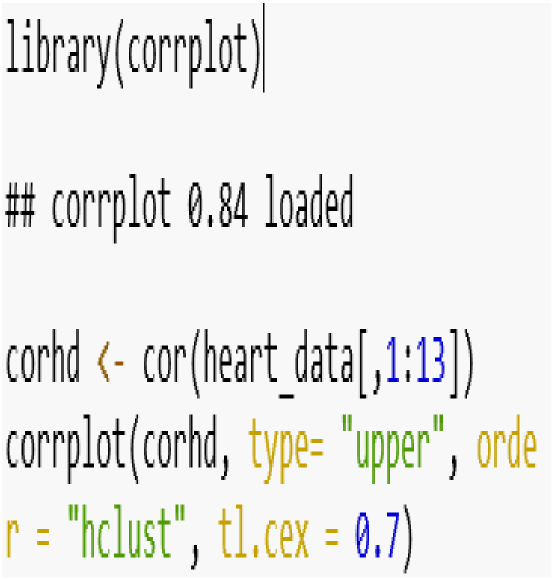
VISUALIZING KEY PREDICTOR/ RISK VARIABLES OF H.D CHOLESTEROL AND AGE



```
##the plot above seeks to find out if any granularity/correlation exists between ag
e and cholesterol levels in either sex. The > ## the plot shows that seems to be th
e case between the ages of 50 - 60-year-old male and female populations. (Juan Bern
al)

> plot (x = heart_data$age, y = heart_data$oldpeak)
```





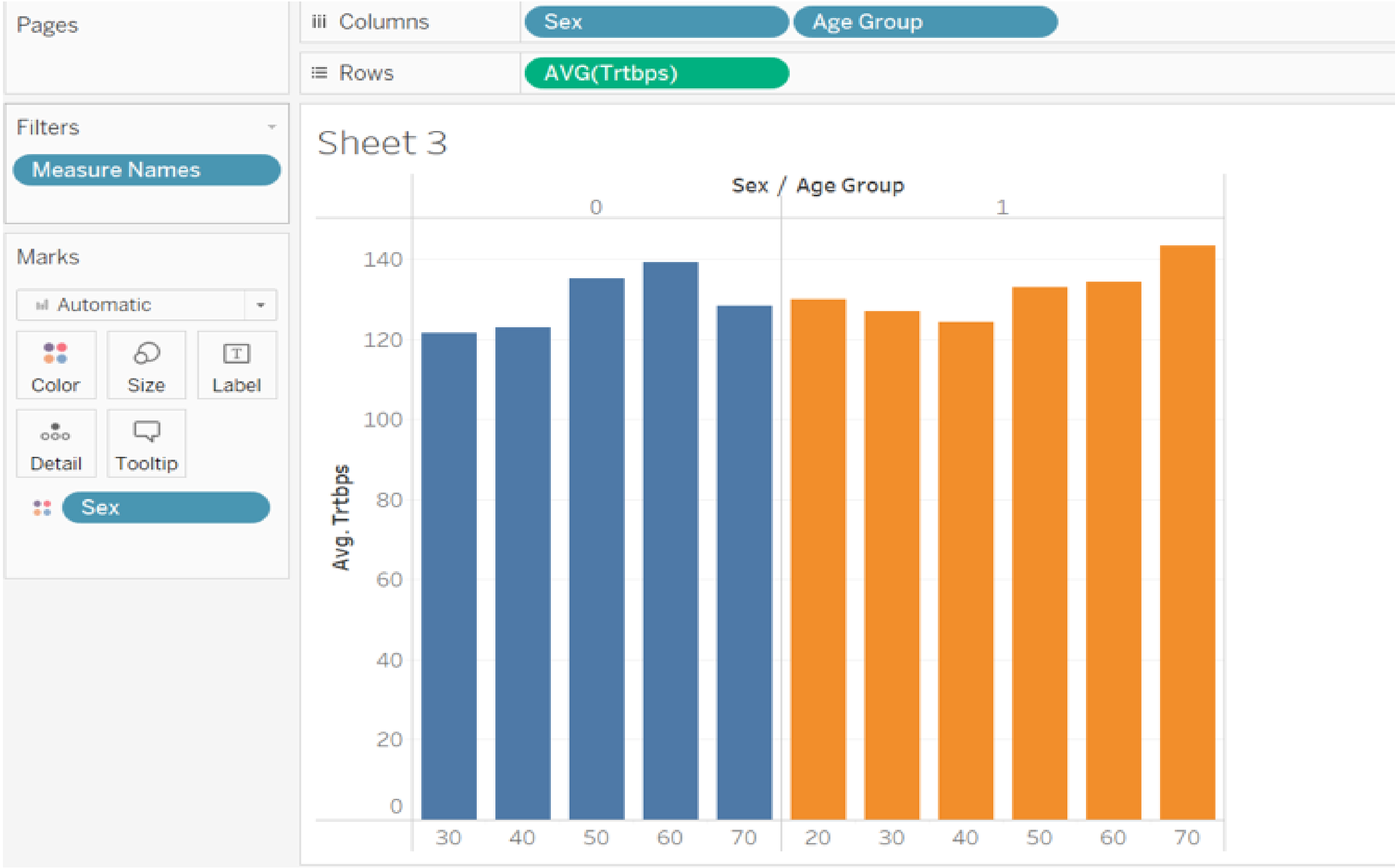
```
library(corrplot)

## corrplot 0.84 loaded

corhd <- cor(heart_data[,1:13])
corrplot(corhd, type= "upper", orde
r = "hclust", tl.cex = 0.7)
```

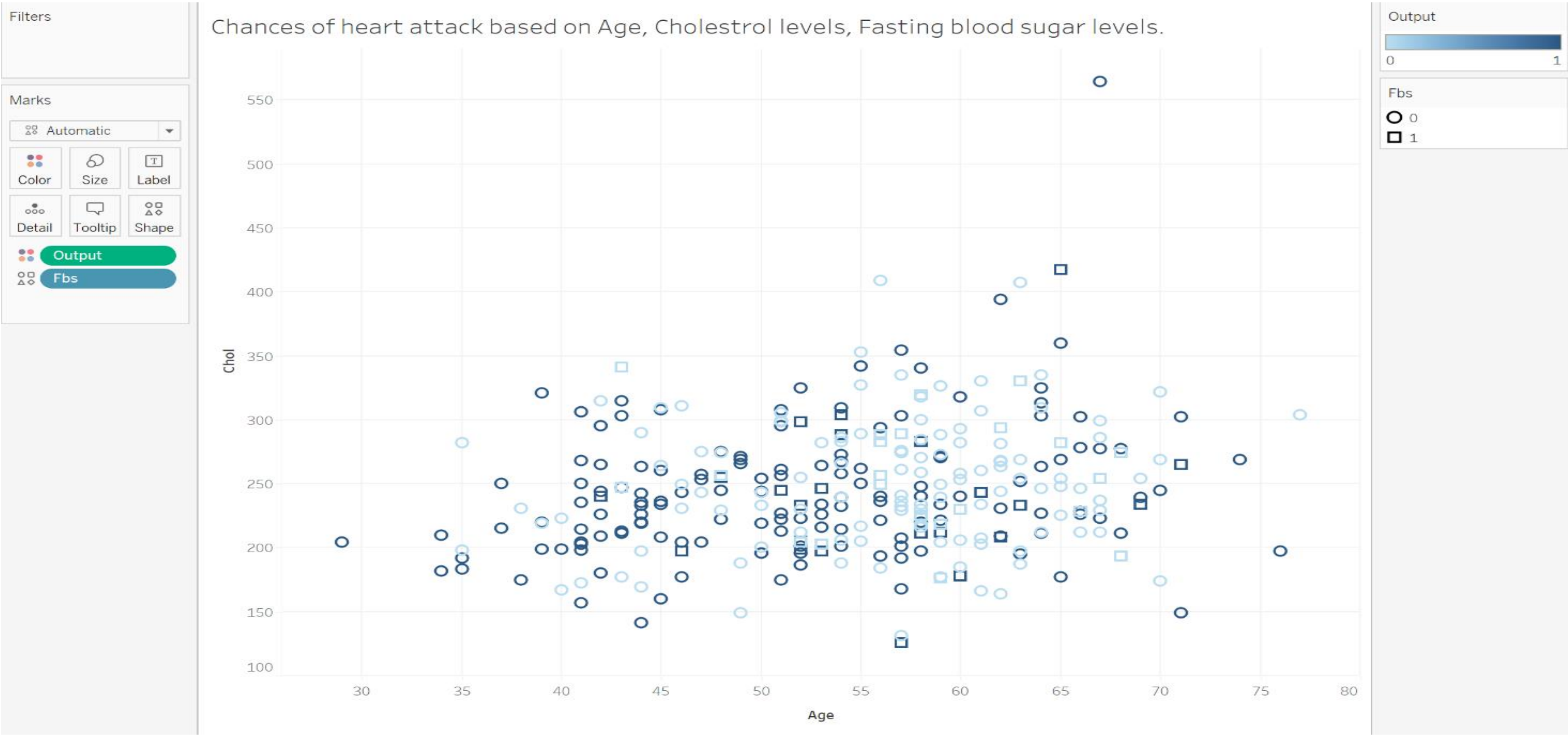
- *It appears to be a strong correlation between **chest pain** and **maximum heart rate reached**, as do the **number of major vessels** and **age**. **Lope** and **old peak*** appear to be inversely correlated.

GENDER, AGE AND BLOOD PRESSURE



The above visualization is a comparision of blood pressure among different age groups of both genders. The bar graph analysis reveals a consistent trend in both genders, indicating that individuals in the age groups between 50-80s tend to have higher blood pressure compared to those in the age groups of 20-50s. The visual representation reinforces the importance of monitoring blood pressure in older age brackets for both men and women

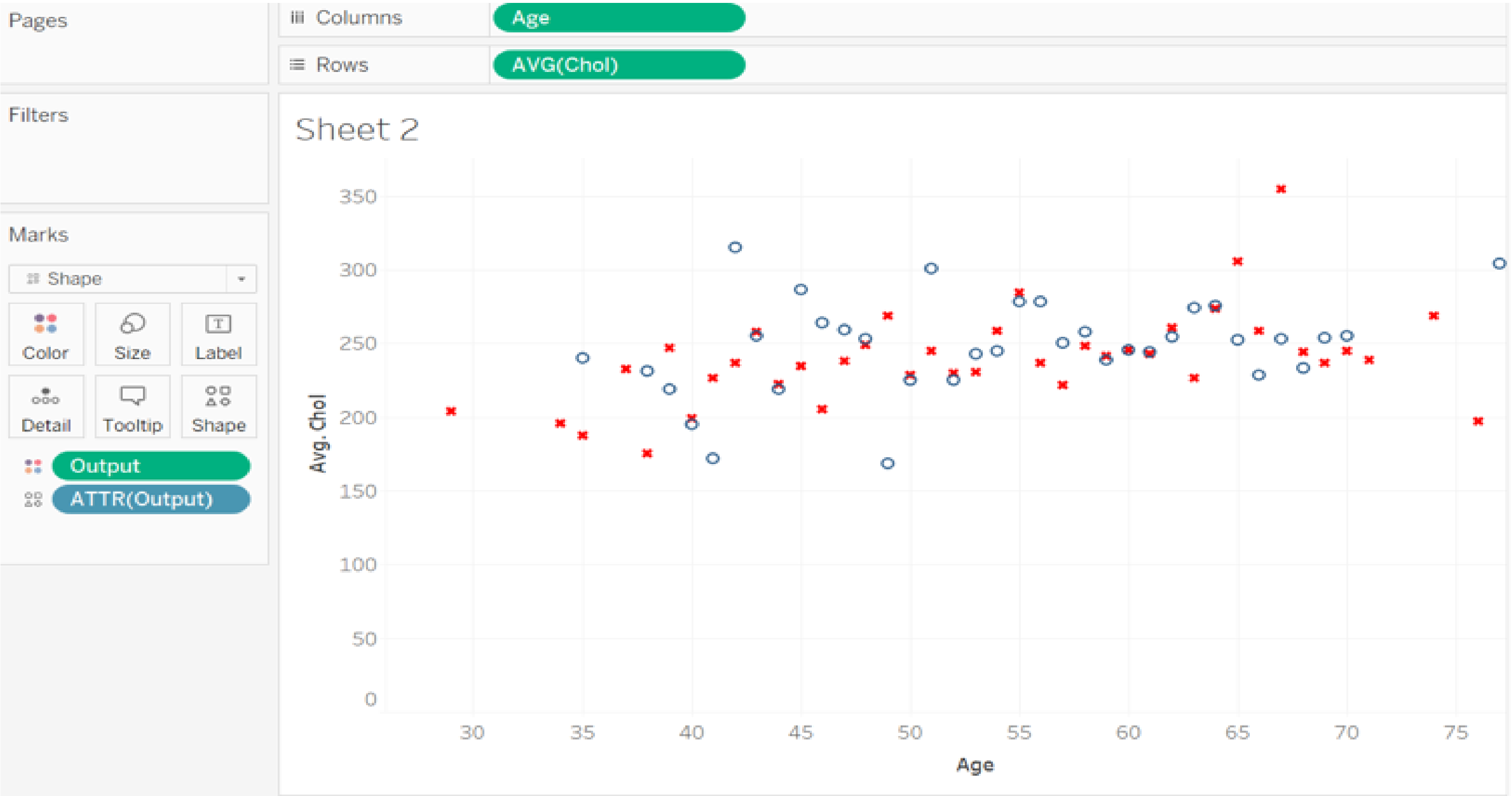
AGE, CHOLESTEROL AND FASTING BLOOD SUGAR LEVELS



The study suggests that a high risk of heart attack occurs when cholesterol levels exceed 200 and fasting blood sugar levels exceed 120. This aligns with clinical information on the link between elevated cholesterol, raised glucose, and cardiovascular risk. This information could aid in preventative measures and healthcare interventions, but individual risk assessments should consider various factors and be conducted in collaboration with healthcare professionals.

AGE, CHOLESTEROL LEVELS AND MORBIDITY FOR BOTH MALES AND FEMALES

Tableau Visualizations



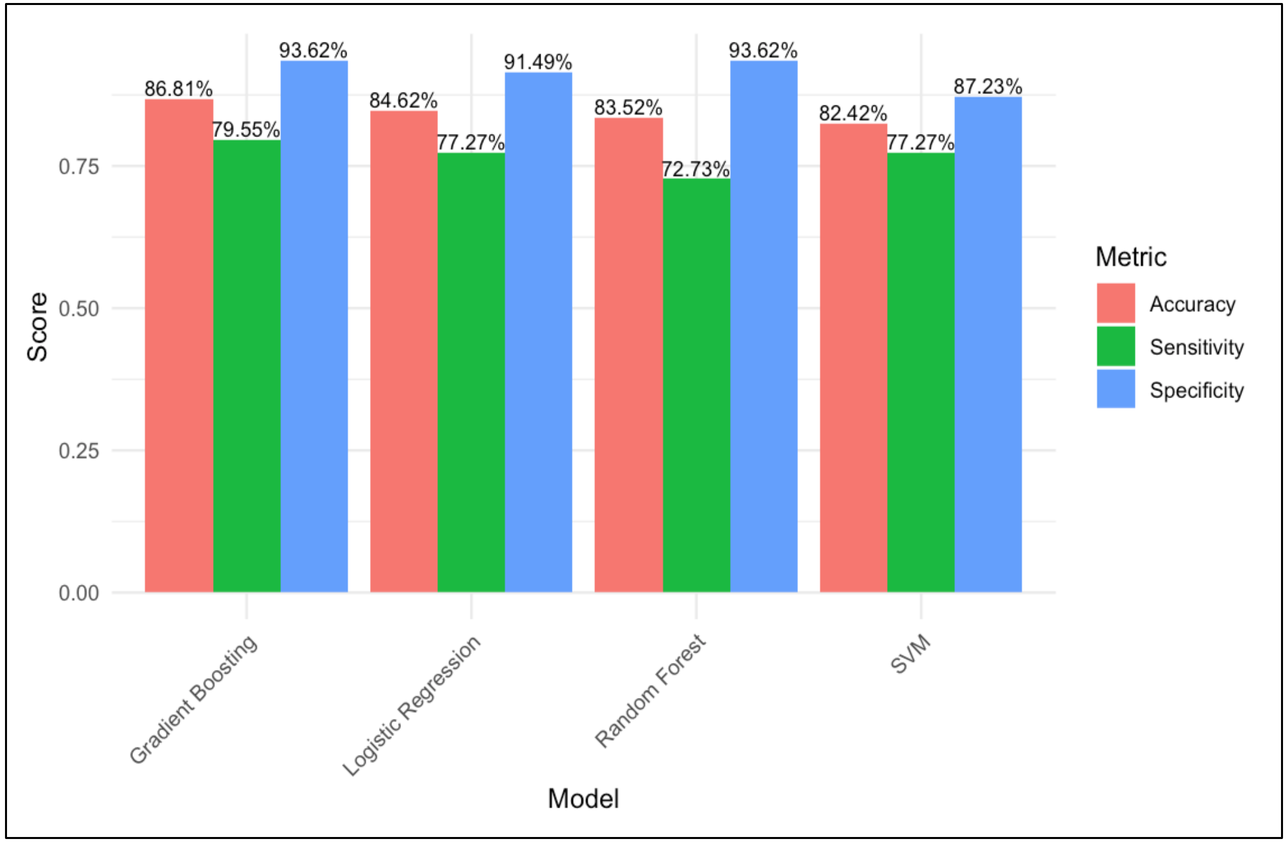
The scatter plot analysis shows that age, cholesterol levels, and cardiovascular failure probability are linked. Red dots indicate heart attacks, while blue dots do not indicate heart attacks. Older individuals with higher cholesterol are more likely to have heart attacks, and vice versa, indicating the importance of cholesterol levels and age in determining the likelihood of heart attacks.

ML PREDICTIVE MODELS

- MODEL PERFORMANCE based in research and in our own findings
- Logistic regression
- Random Forest
- Gradient Boosting
- Support Vectors
- Area under the curve ROC
- The Independent or predictive variables in a Machine Learning Model are independent of each other. The outcome and performance of the model depends on a set of variables that are not related to each other. It calculates the probability of an event occurring based on information about prior events.

CONFUSION MATRIX

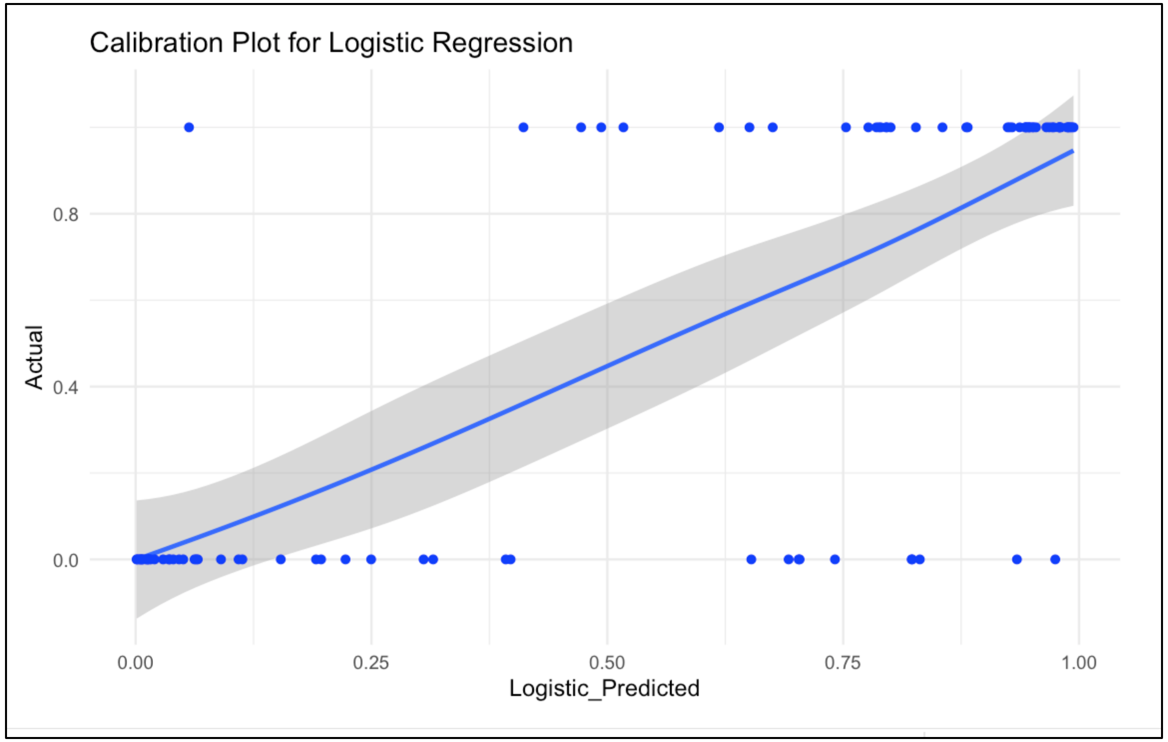
- Gradient Boosting has the tallest red bar, indicating it has the highest overall accuracy. Its green bar, representing Sensitivity, is also quite high, though not the highest, suggesting it's quite adept at identifying patients with heart failure. Notably, its blue bar for Specificity matches that of the Random Forest model, both scoring very well, indicating a strong ability to recognize patients without the condition.
- The Logistic Regression model shows strong performance across all metrics, with relatively balanced bars. The Random Forest model, while slightly lower in accuracy, exhibits exceptional Specificity. Lastly, the SVM has competitive scores but falls behind particularly in Specificity.



PREDICTIVE MODELS VISUALIZATIONS

- Various machine learning models have been applied, including Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM). For model validation, we used accuracy as the metric and have employed ROC curves and calibration plots to evaluate the performance of the models.

- Finally, the accuracy of the models is listed, with Gradient Boosting performing the best at 86.8%, followed by Logistic Regression at 84.6%, Random Forest at 83.5%, and SVM at 82.4%.

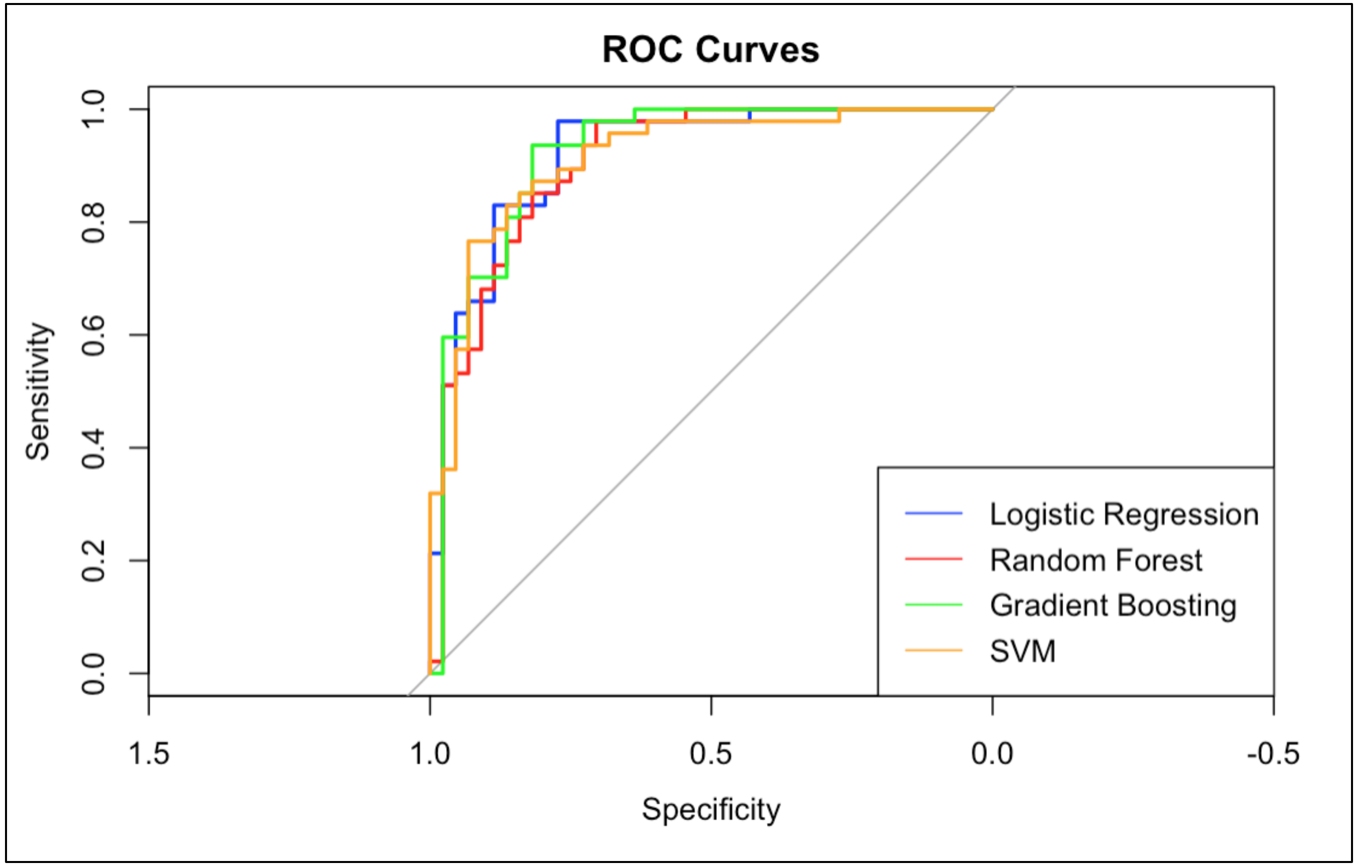


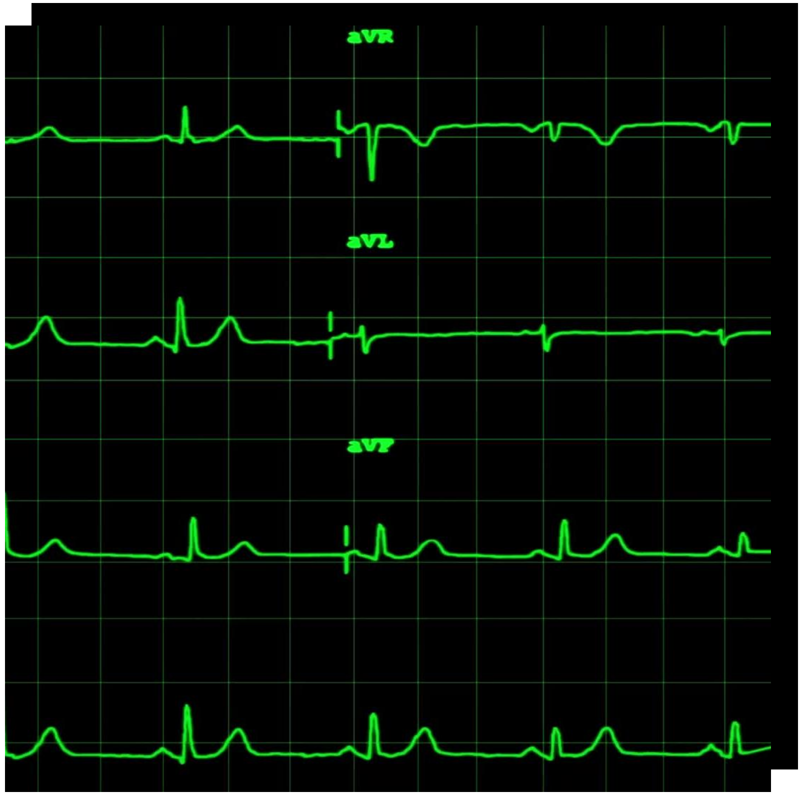
BEST PERFORMING PREDICTIVE MODELS

- From the ROC curves, it's visible that all models perform significantly better than random guessing, with the area under the curve (AUC) being substantially above 0.5 for all.

- After comparing the models, the Gradient Boosting model stands out with the highest AUC, reflecting its superior ability to distinguish between patients with and without heart failure. Its higher accuracy further solidifies its position as the best model for our project. This means that in a clinical setting, Gradient Boosting is more likely to correctly identify patients at risk, which is paramount for early intervention and treatment planning.

- In summary, our evaluation based on both the calibration plot and the ROC chart suggests that the Gradient Boosting model is the best predictor of heart failure in our dataset.





HEART DISEASE KEY RISK FACTORS IDENTIFIED

THE SIX(6) MAIN CULPRITS:

- Age
- Gender
- high blood pressure
- high cholesterol
- Smoking and diabetes

CONCLUSIONS AND FINDINGS
CARDIOVASCULAR DISEASE PREDICTION (CVD)

- **Most significant correlations** (Gender, Age, Blood Pressure and to some extent cholesterol levels)
- Men and women over 45 might be less inclined to follow exercise routines and healthier nutritional habits
- **165/303= 55 % of the 303 patients i**n the dataset are more prone to develop CAV.
- **Age findings** Men between 49 – 59 are more prone to CAV. Women > 70
- Key **Relevant research findings** identify risk factors that show similarities with our findings
- (Age, Gender, BP and cholesterol levels are the main risk factors in CAV across studies.
- **Statistical relevance** – The ROC Area under the Curve above 90% shows the diagnostic ability of the binary classification True Positive Rate vs False Positive rate. The closer to a 90 degree angle the higher the accuracy of the classification prediction.
- **The accuracy of the models** is listed, with Gradient Boosting performing the best at 86.8%, followed by Logistic Regression at 84.6%, Random Forest at 83.5%, and SVM at 82.4%.
- **The Gradient Boosting model** stands out with the highest AUC, reflecting its superior ability to distinguish between patients with and without heart failure.
- **The Logistic Regression model** shows strong performance across all metrics, with relatively balanced bars. The Random Forest model, while slightly lower in accuracy, exhibits exceptional Specificity. Lastly, the SVM has competitive scores but falls behind particularly in Specificity.
- **Predictive modeling** and decision making is based on the accuracy of the models and that statistics significance of sample
- Stakeholder impact takes a holistic approach on how all actors involved in patient care are key in preventing CAD

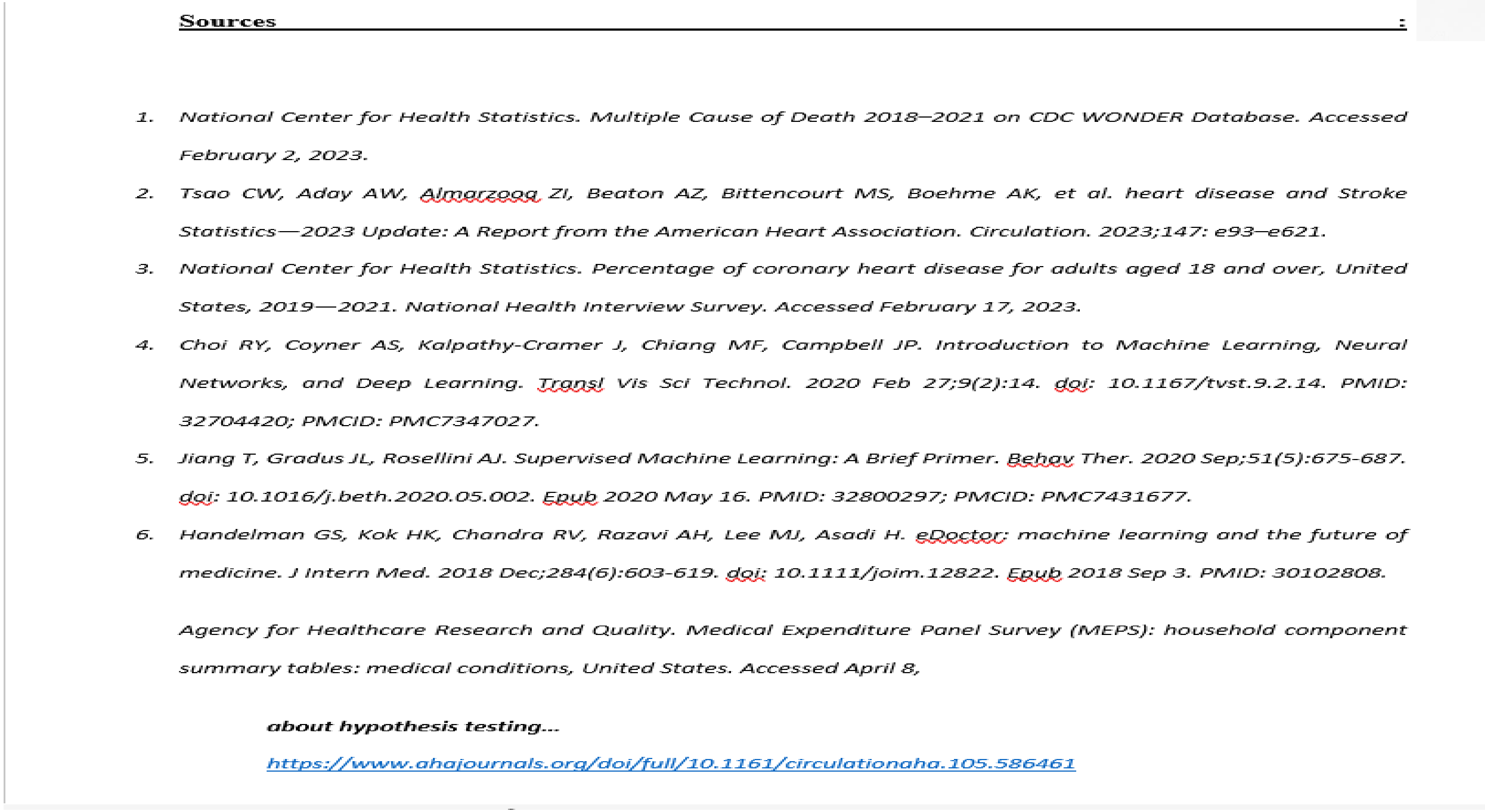
CONTINUATION CONCLUSIONS

- Age propensity to heart attack women .> 70
- Age propensity heart attack women <50-60>
- Data-driven customer-centric analytics tools such as the ones we implemented in our project allow for objective decision making and as preventive mechanisms against heart disease.
- Stakeholder involvement (Doctors, Nurses, clinical personnel, SW, community advocates , Insurance companies, extended family support and proper training and patient education are key health forces that will minimize the incidence of CVD in both male and female population.

THE KEY IS ALWAYS YOU!

THANK YOU VERY MUCH !

REFERENCES



Sources :

1. *National Center for Health Statistics. Multiple Cause of Death 2018–2021 on CDC WONDER Database. Accessed February 2, 2023.*
2. *Tsao CW, Aday AW, Almarzooq ZI, Beaton AZ, Bittencourt MS, Boehme AK, et al. heart disease and Stroke Statistics—2023 Update: A Report from the American Heart Association. Circulation. 2023;147: e93–e621.*
3. *National Center for Health Statistics. Percentage of coronary heart disease for adults aged 18 and over, United States, 2019—2021. National Health Interview Survey. Accessed February 17, 2023.*
4. *Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020 Feb 27;9(2):14. doi: 10.1167/tvst.9.2.14. PMID: 32704420; PMCID: PMC7347027.*
5. *Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. Behav Ther. 2020 Sep;51(5):675-687. doi: 10.1016/j.beth.2020.05.002. Epub 2020 May 16. PMID: 32800297; PMCID: PMC7431677.*
6. *Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018 Dec;284(6):603-619. doi: 10.1111/joim.12822. Epub 2018 Sep 3. PMID: 30102808.*

Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey (MEPS): household component summary tables: medical conditions, United States. Accessed April 8,

about hypothesis testing…

https://www.ahajournals.org/doi/full/10.1161/circulationaha.105.586461