

# KAGGLE CARDIAC DATA

## DATA EXPLORATION AND ANALYSIS

SAM LOYD  
NOVEMBER 2019

# KAGGLE CARDIAC DATA CODE BOOK

PROVIDED BY SVETLANA UNLIANOVA AT RYERSON UNIVERSITY

[HTTPS://WWW.KAGGLE.COM/SULIANOVA/EDA-CARDIOVASCULAR-DATA/NOTEBOOK#EDA-OF-CADIOVASCULAR-DISEASES-DATA](https://www.kaggle.com/sulianova/eda-cardiovascular-data/notebook#eda-of-cardiovascular-diseases-data)

## NUMERICAL DATA

AGE - days

HEIGHT - cm

WEIGHT - kg

SYSTOLIC BLOOD PRESSURE

DIASTOLIC BLOOD PRESSURE

## BINARY DATA

GENDER

SMOKING

ALCOHOL

PHYSICAL ACTIVITY

**CARDIOVASCULAR DISEASE**

## CATEGORICAL DATA

CHOLESTEROL

GLUCOSE

## HYPOTHESIS

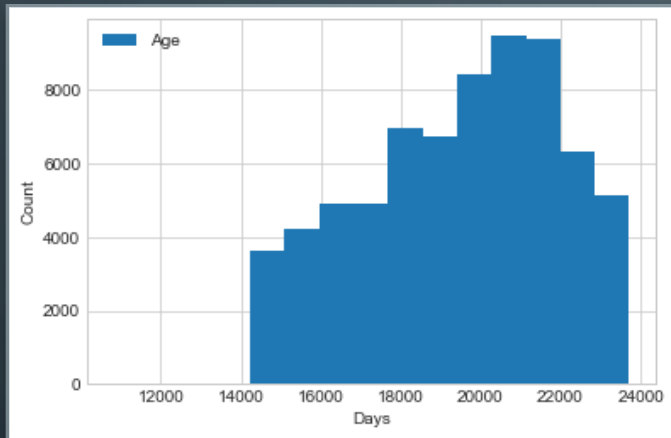
“

DOES BEING OVERWEIGHT AS MEASURED BY THE STANDARD BMI FORMULA USING WEIGHT AND HEIGHT, TEND TO INCREASE THE LIKELIHOOD OF HAVING A CARDIOVASCULAR DISEASE?

”

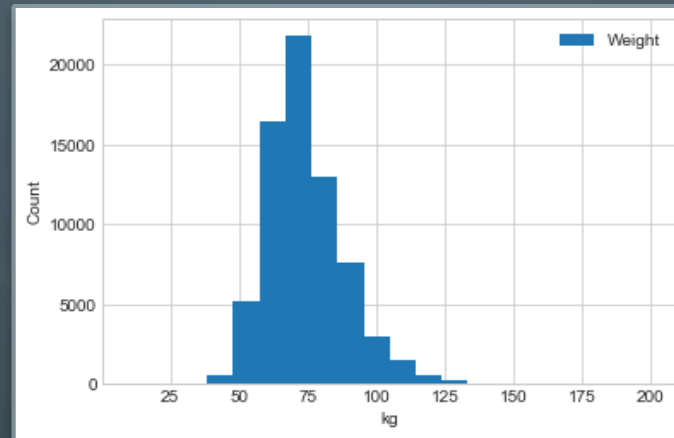
How does gender impact the answer to the question above if at all?

# EXAMPLES FROM DISTRIBUTION ANALYSIS



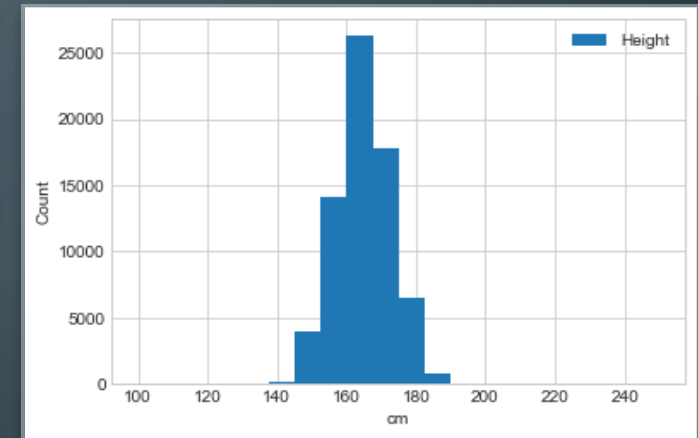
## AGE

Statistics indicate symmetry and platykurtic (light tailed).



## WEIGHT

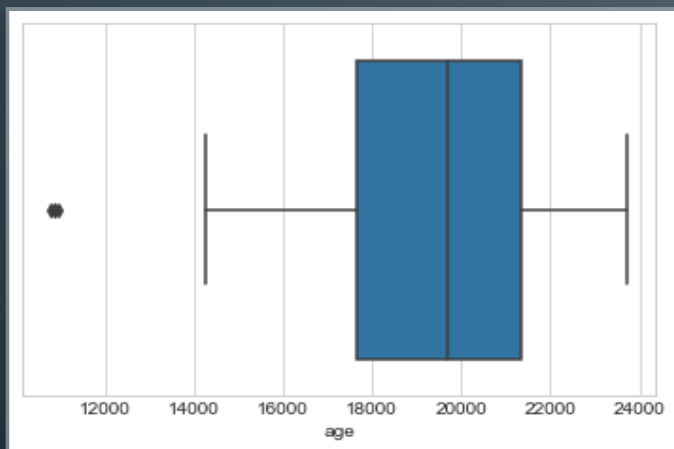
Statistics show positive skew and leptokurtic (heavy tailed).



## HEIGHT

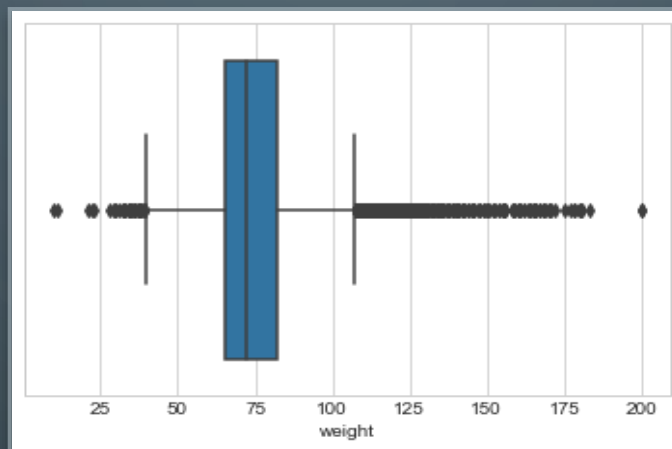
Statistics indicate negative skew and leptokurtic.

# EXAMPLES FROM OUTLIER ANALYSIS



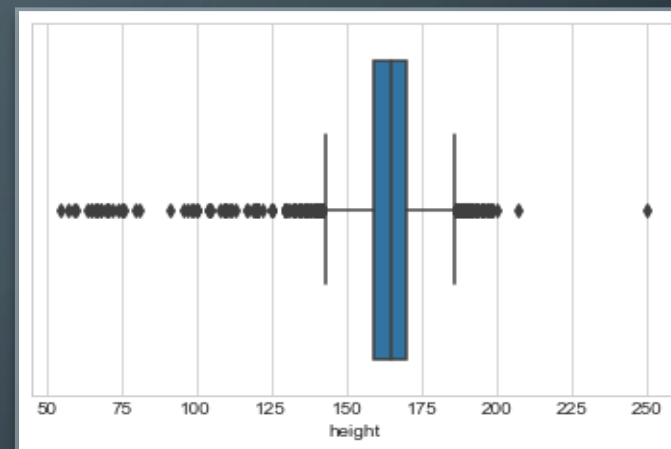
## AGE

Age 29 was reasonable so not removed.



## WEIGHT

Removed cases over 181 kg and less than 36 kg.

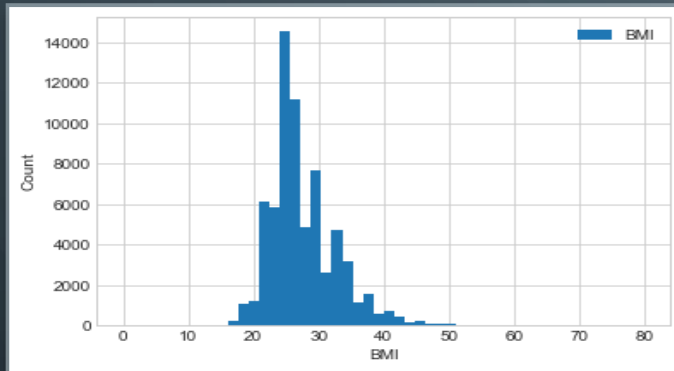


## HEIGHT

Removed top outlier to the right and all less than 121 cm or 4 ft.

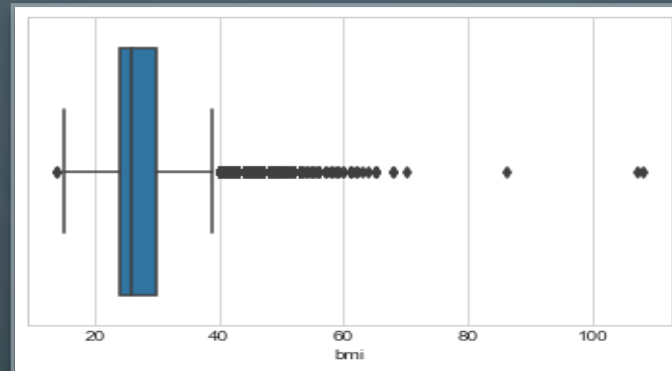
# CALCULATED BMI

The formula for **BMI** is weight in kilograms divided by height in meters squared. Using this data set, that is weight divided by the square of the height after it is divided by 100.



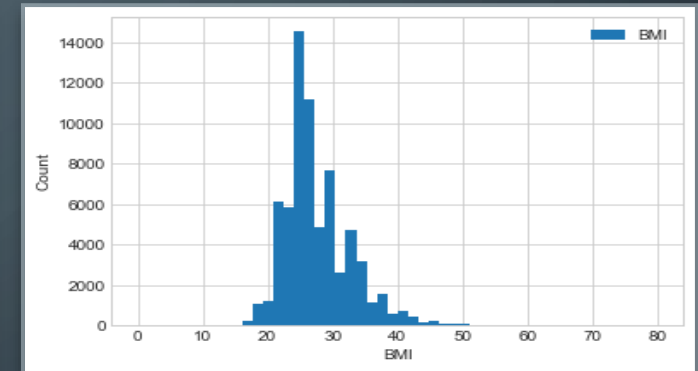
## INITIAL HISTOGRAM

Positive skew and leptokurtotic.



## BOX PLOT - OUTLIERS

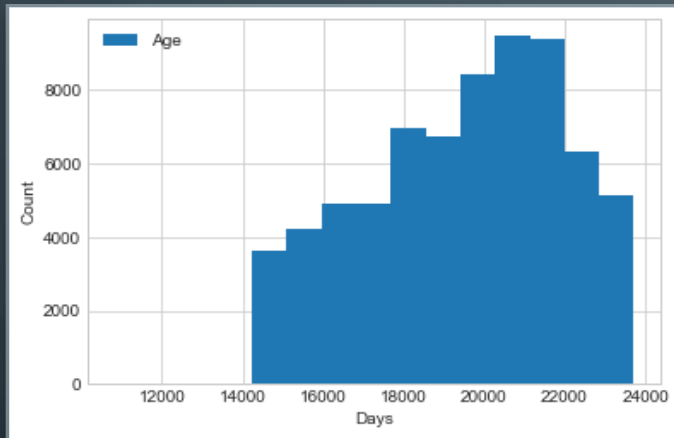
Removed extreme cases based on domain knowledge.



## FINAL HISTOGRAM

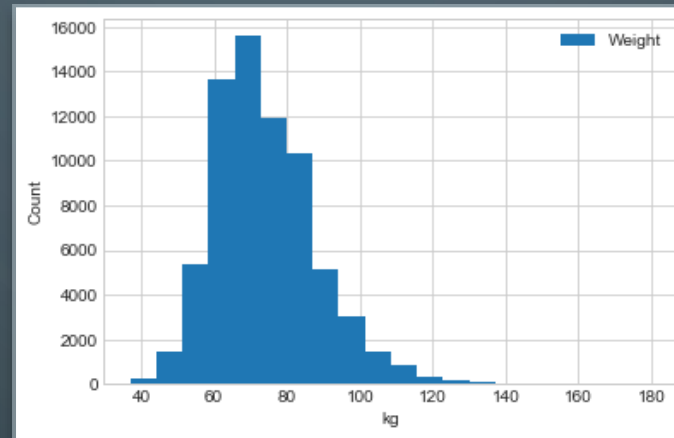
Slight kurtosis improvement which was hard to detect visually.

# FOLLOW UP ANALYSIS POST OUTLIER REMOVAL



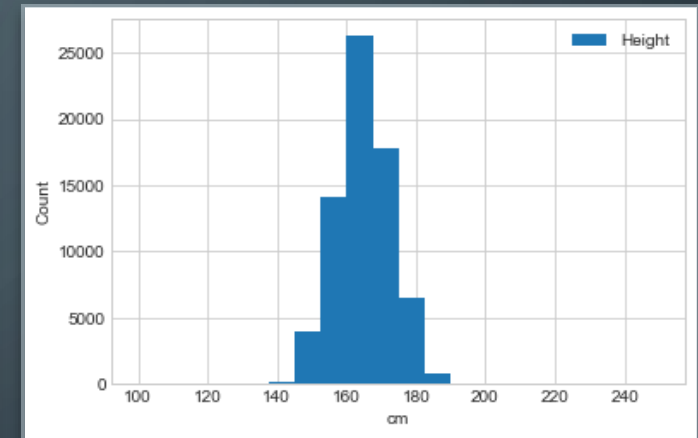
## AGE

No changes to outlier or analysis.



## WEIGHT

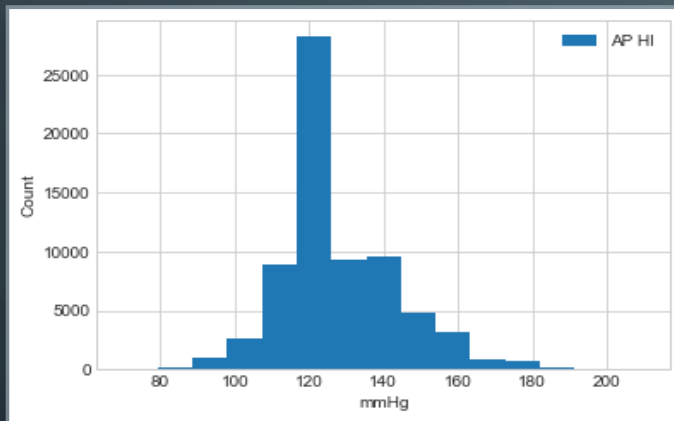
Lessened kurtosis.



## HEIGHT

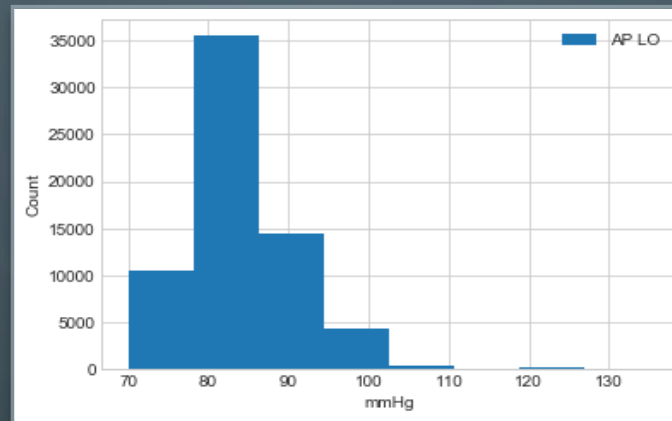
Most improved example with relatively normal statistics.

# CONTINUED ANALYSIS POST OUTLIER REMOVAL



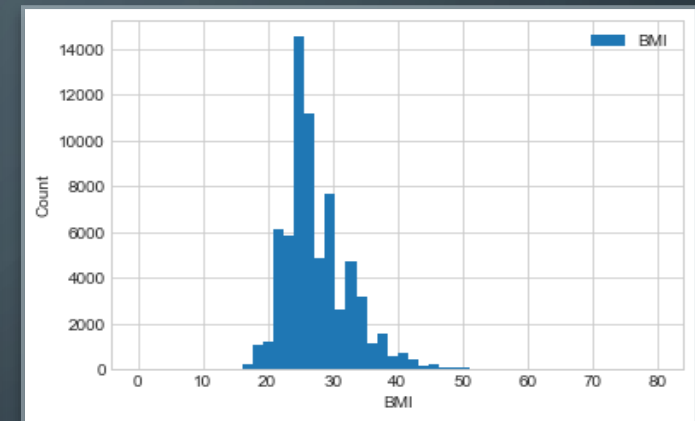
## SYSTOLIC

Slight positive skew and leptokurtic.



## DIASTOLIC

Relatively normal skew and leptokurtic.

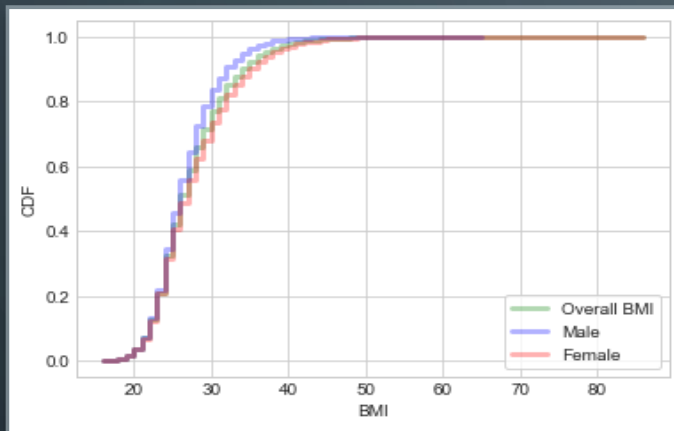


## BMI

Positive skew and leptokurtic.

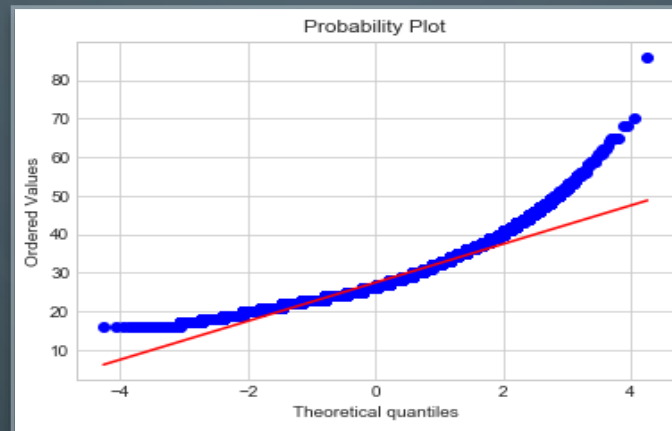


# FURTHER EVIDENCE OF NON GAUSSIAN DISTRIBUTION FOR BMI



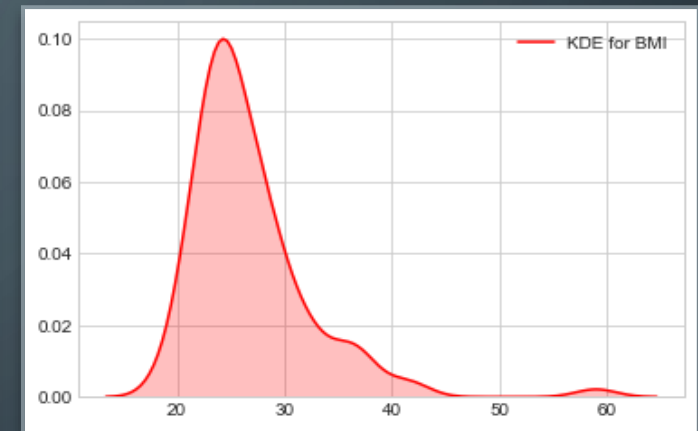
## CDF

Plot indicates skew.



## PROBABILITY PLOT

Note the deviations from the line.

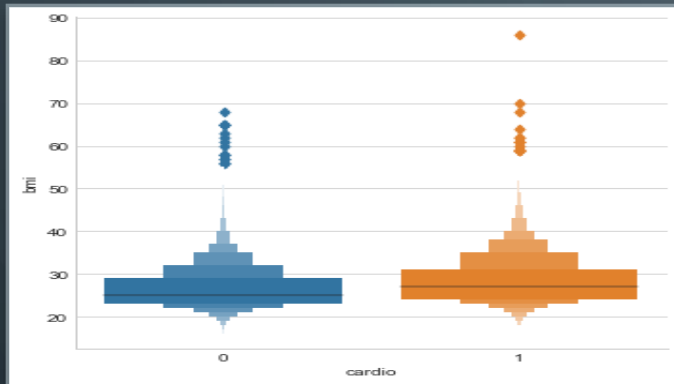


## KDE

Non-symmetrical indicating skew.

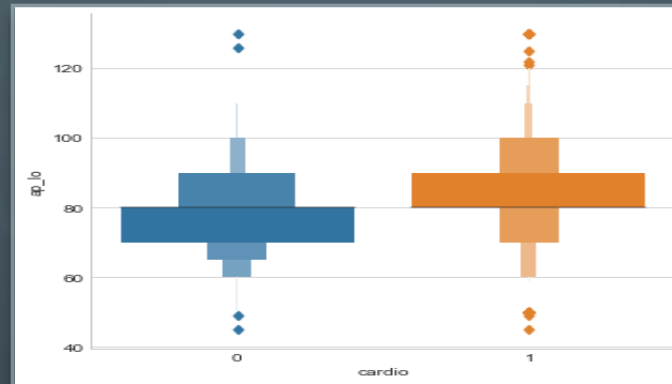
# CARDIO VISUALIZATIONS USING BOXEN PLOTS

Due to the binary nature of having a cardio disease used in this analysis, enhanced box plots from the python based seaborn package were utilized for visualizations.



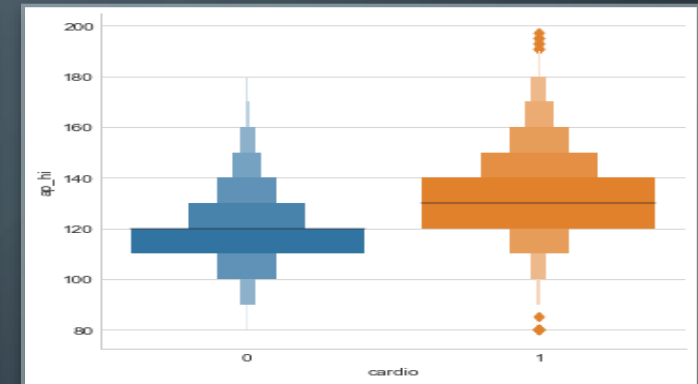
## BMI - CARDIO

Correlation appears slightly noticeable.



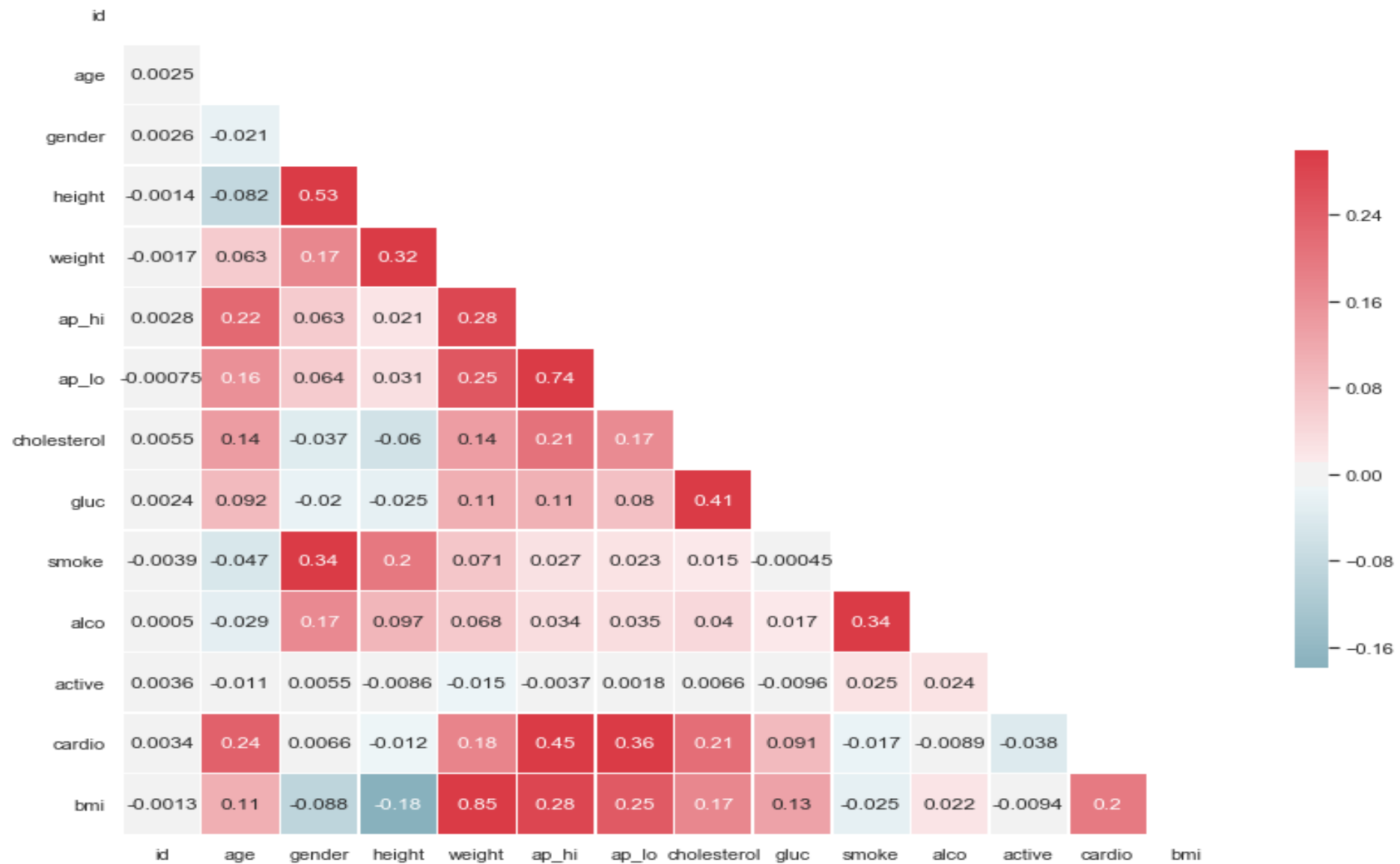
## DIASTOLIC - CARDIO

Indicates some positive correlation.

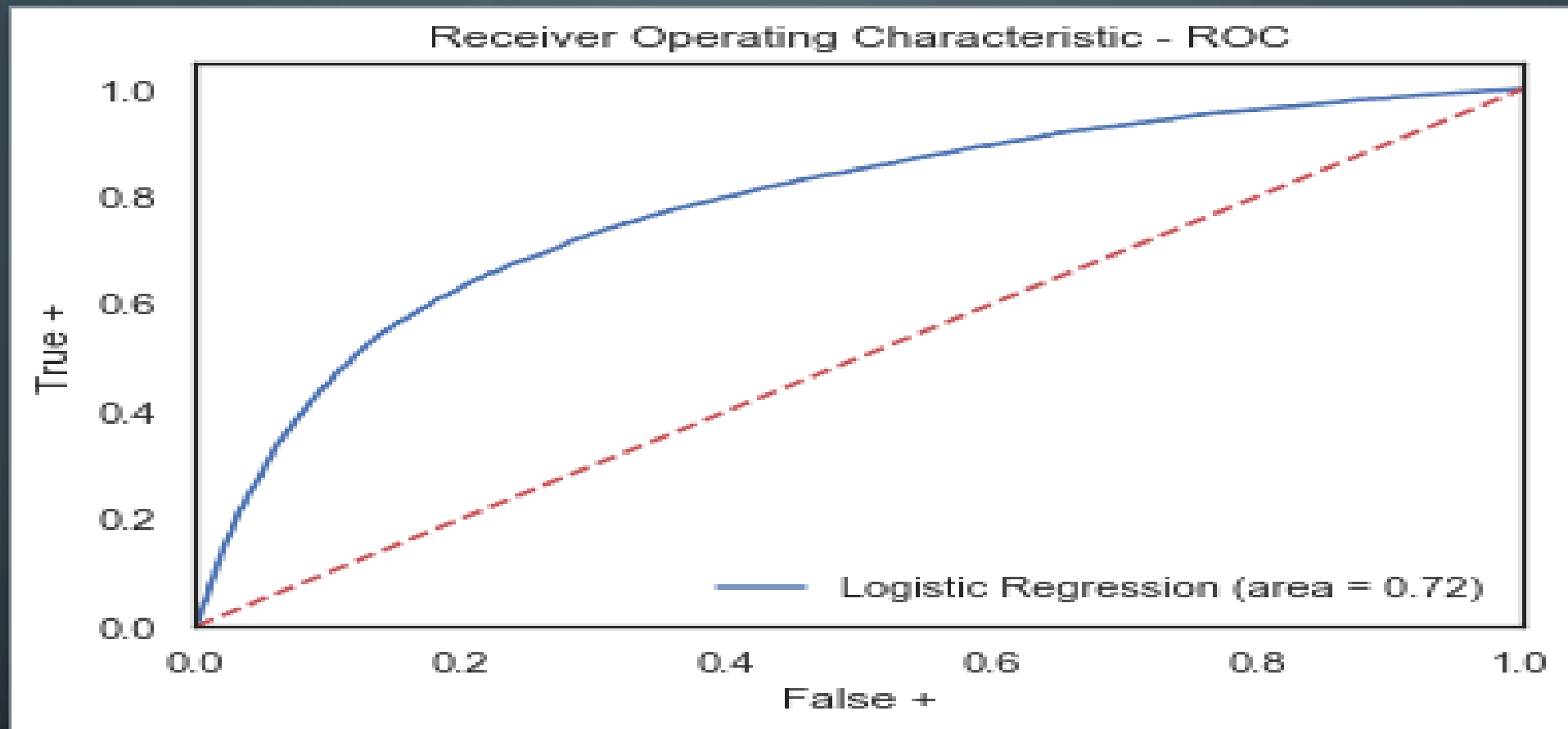


## SYSTOLIC - CARDIO

Strongest visual of positive correlation.



Correlation Heatmap Using Spearman's Method



ROC Curve from Sklearn's LogisticRegression Function

## CONCLUSION

“

CORRELATION BETWEEN BMI AND CARDIOVASCULAR DISEASE, WHILE STATISTICALLY SIGNIFICANT, WAS WEAK. SYSTOLIC BLOOD PRESSURE HAD A STRONGER CORRELATION AND WAS MORE USEFUL IN CONSTRUCTING PREDICTIVE MODELS.

”

Gender as confirmed by multiple tests did not prove to have a correlation with heart disease in this population. Given the unknown amount of sampling and other bias in this population set, I would discourage projecting this conclusion onto a larger population set such as the general public.