KAGGLE CARDIAC DATA
DSC 530
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

NUMERICAL DATA

AGE - days

HEIGHT - cm

WEIGHT - kg

SYSTOLIC BLOOD PRESSURE

DIASTOLIC BLOOD PRESSURE

BINARY DATA

GENDER

SMOKING

ALCOHOL

PHYSICAL ACTIVITY

<u>CARDIOVASCULAR DISEASE</u>

CATEGORICAL DATA

CHOLESTEROL

GLUCOSE

HYPOTHESIS

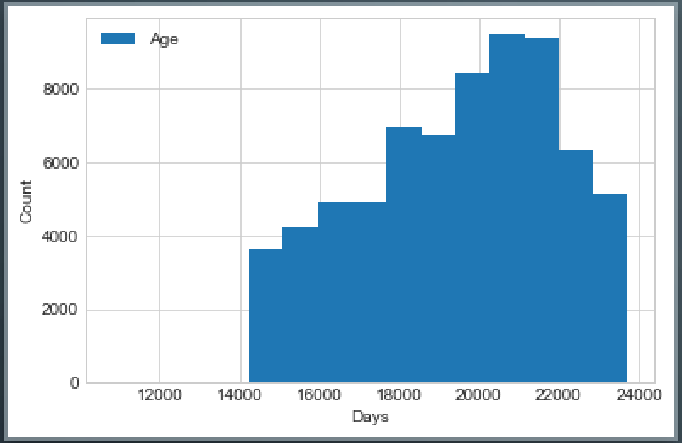
“

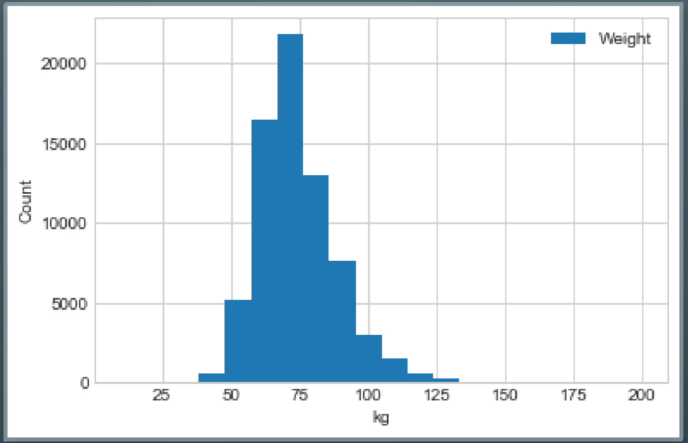
DOES BEING OVERWEIGHT AS MEASURED BY THE STANDARD BMI FORMULA USING WEIGHT AND HEIGHT, TEND TO INCREASE THE LIKELIHOOD OF HAVING A CARDIOVASCULAR DISEASE?

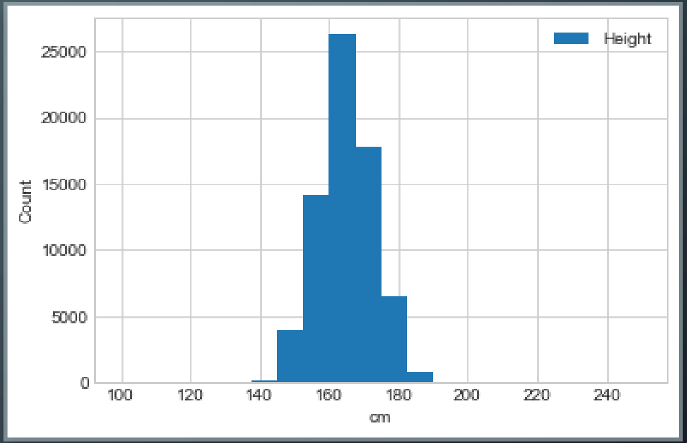
”

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).

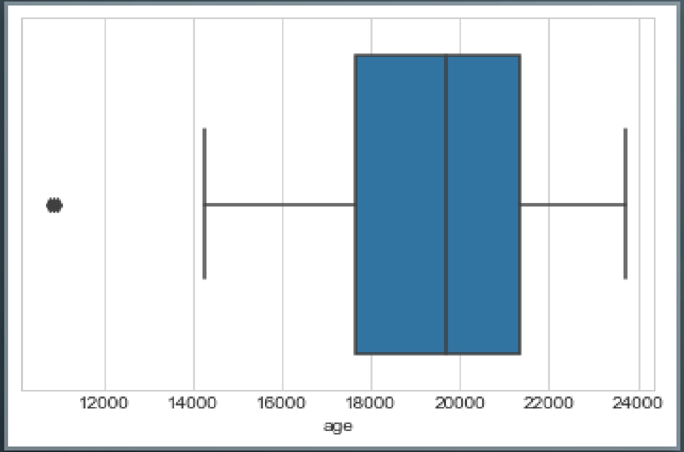
WEIGHT

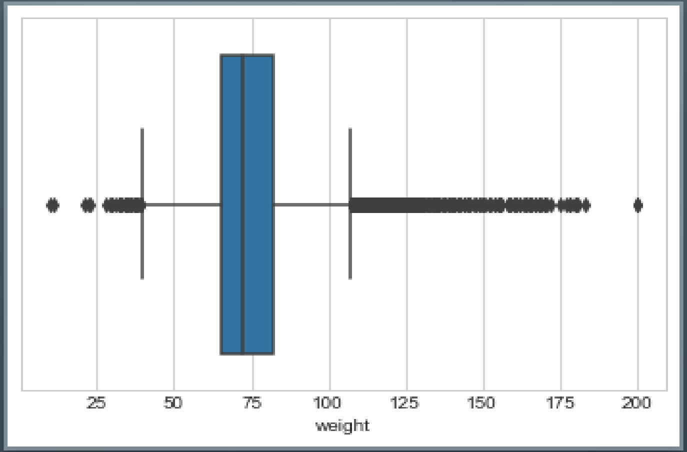
Statistics show positive skew and leptokurtic (heavy tailed).

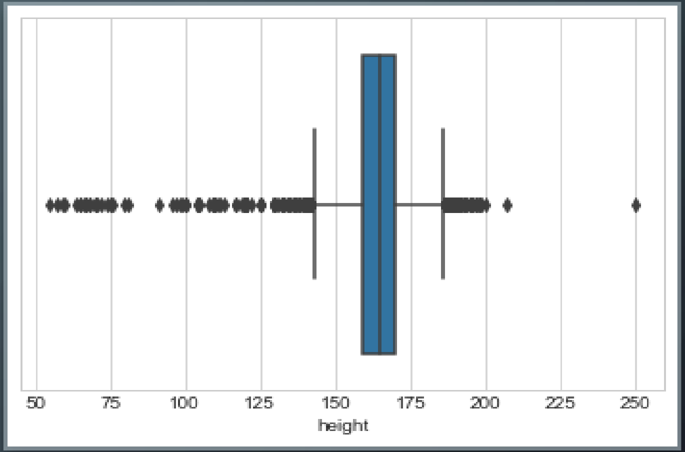
HEIGHT

Statistics indicate negative skew and leptokurtic.

EXAMPLES FROM OUTLIER ANALYSIS







AGE

Age 29 was reasonable so not removed.

WEIGHT

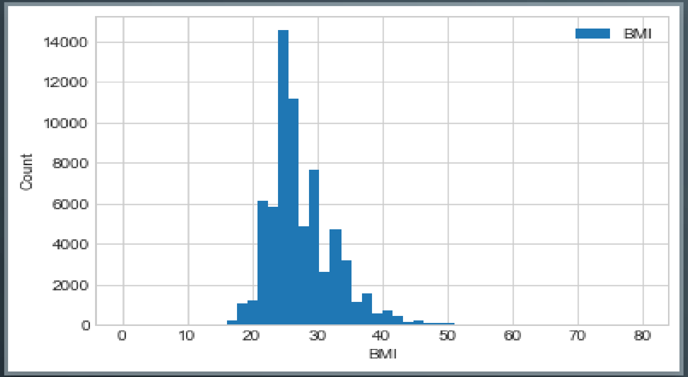
Removed cases over 181 kg and less than 36 kg.

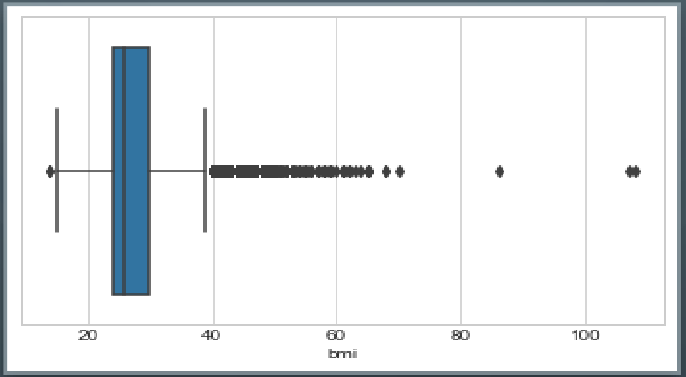
HEIGHT

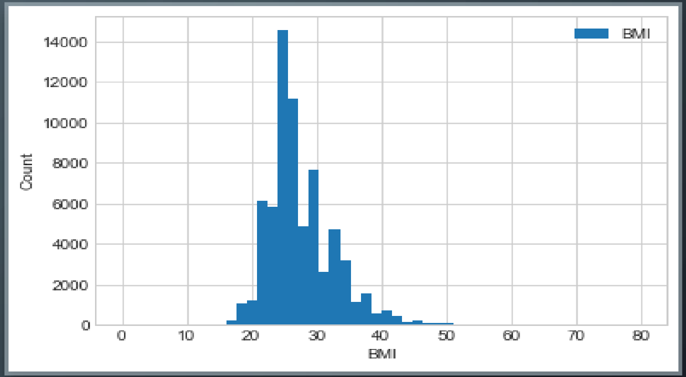
Removed top outlier to the right and all less than 121 cm or 4 ft.

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.

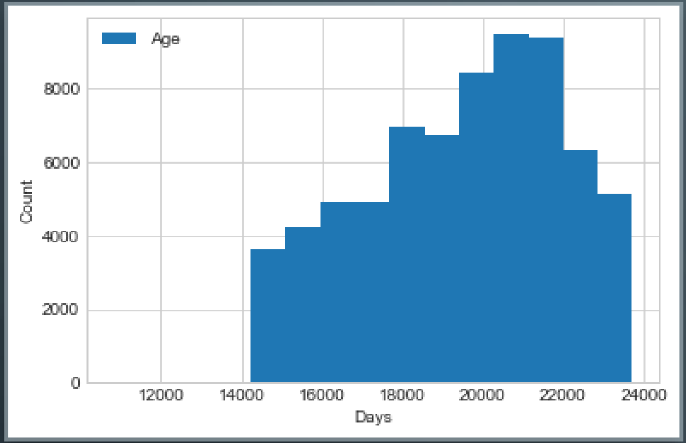
BOX PLOT - OUTLIERS

Removed extreme cases based on domain knowledge.

FINAL HISTOGRAM

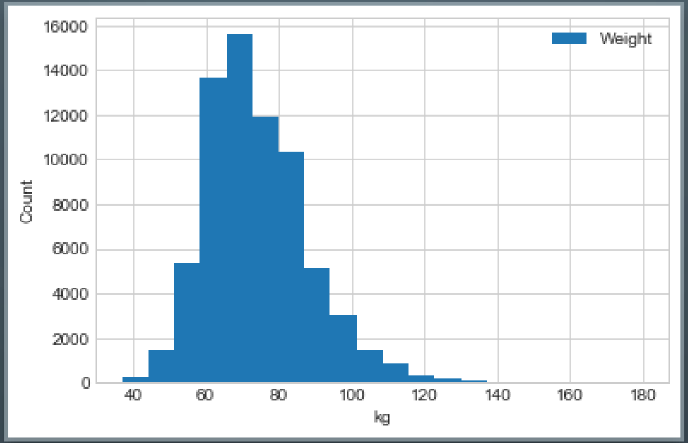
Slight kurtosis improvement which was hard to detect visually.

FOLLOW UP ANALYSIS POST OUTLIER REMOVAL



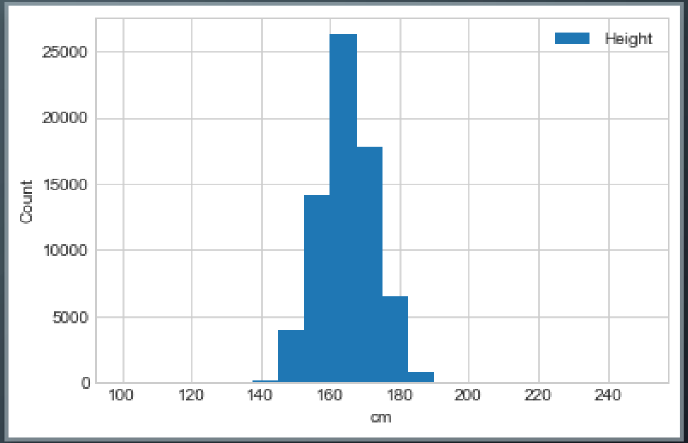
AGE

No changes to outlier or analysis.



WEIGHT

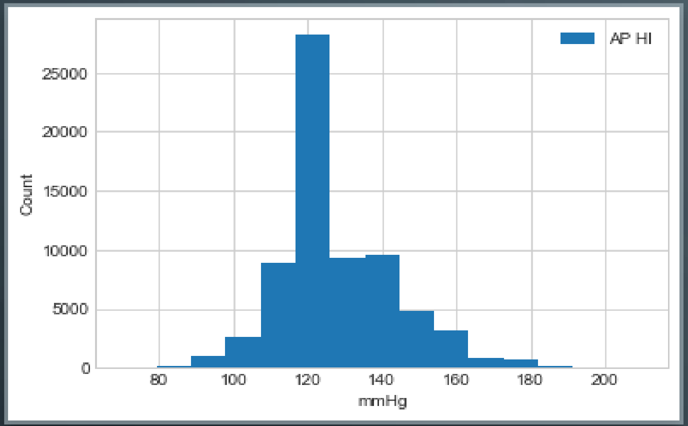
Lessened kurtosis.



HEIGHT

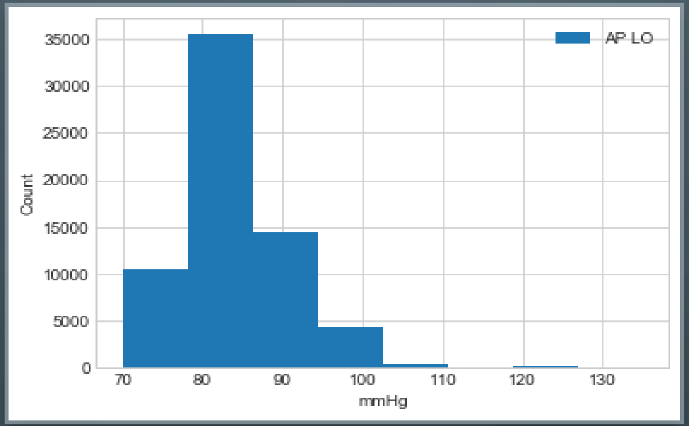
Most improved example with relatively normal statistics.

CONTINUED ANALYSIS POST OUTLIER REMOVAL



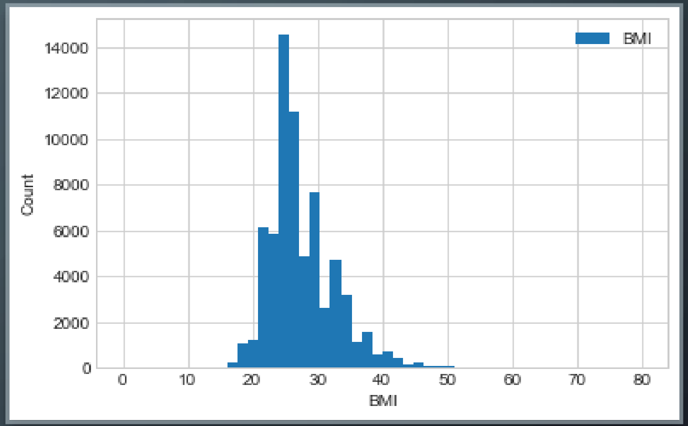
SYSTOLIC

Slight positive skew and leptokurtic.



DIASTOLIC

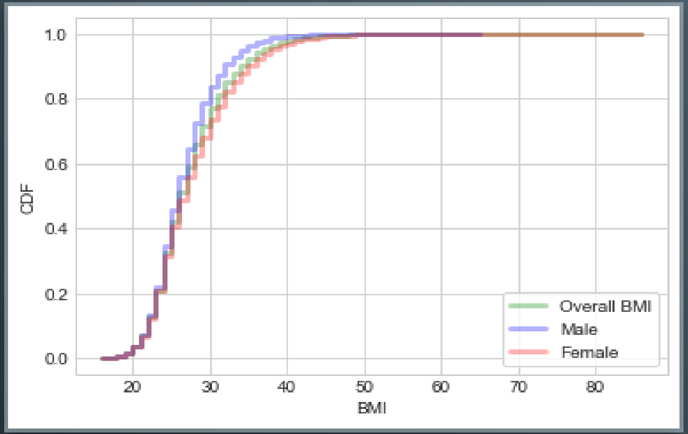
Relatively normal skew and leptokurtic.



BMI

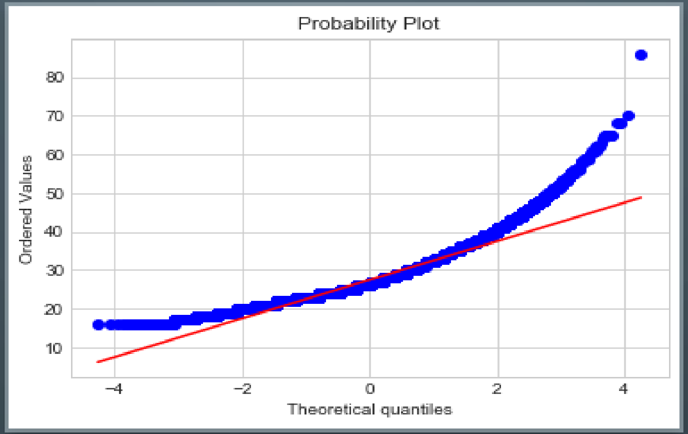
Positive skew and leptokurtic.

FURTHER EVIDENCE OF NON GAUSSIAN DISTRIBUTION FOR BMI



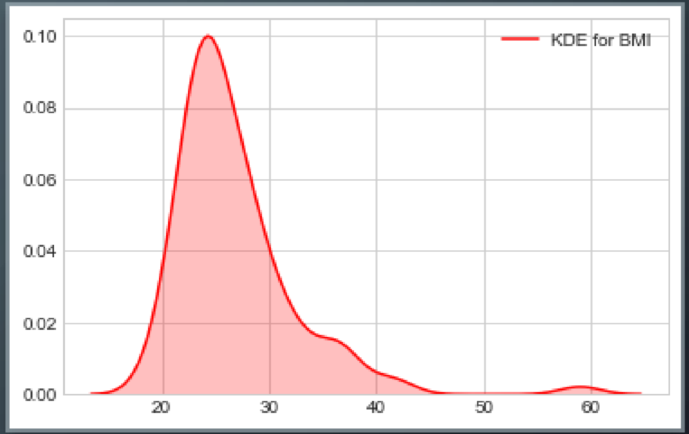
CDF

Plot indicates skew.



PROBABILITY PLOT

Note the deviations from the line.

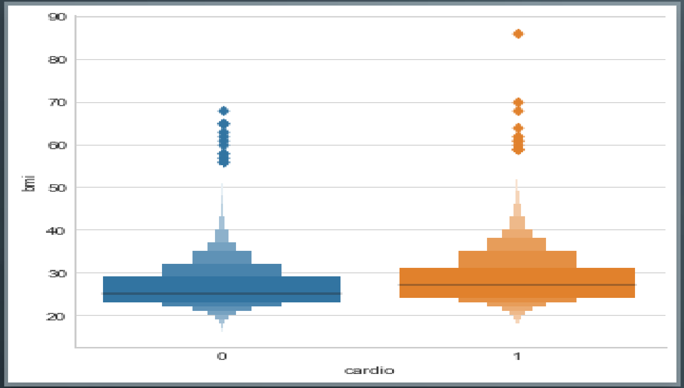


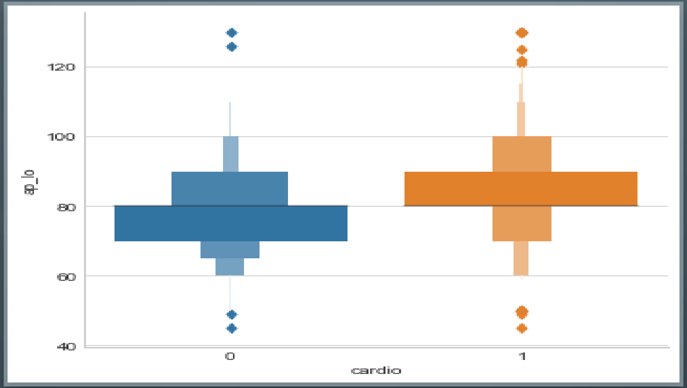
KDE

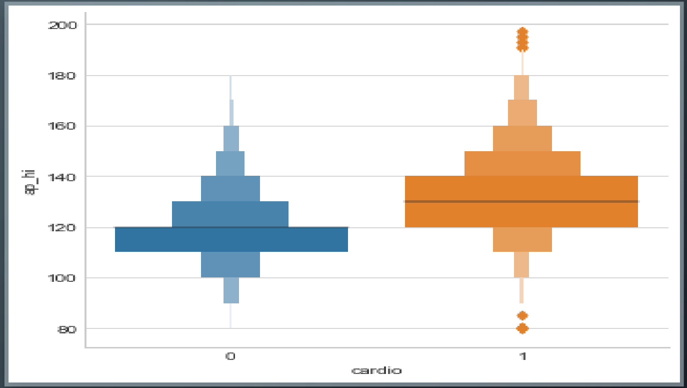
Non-symmetrical indicating skew.

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

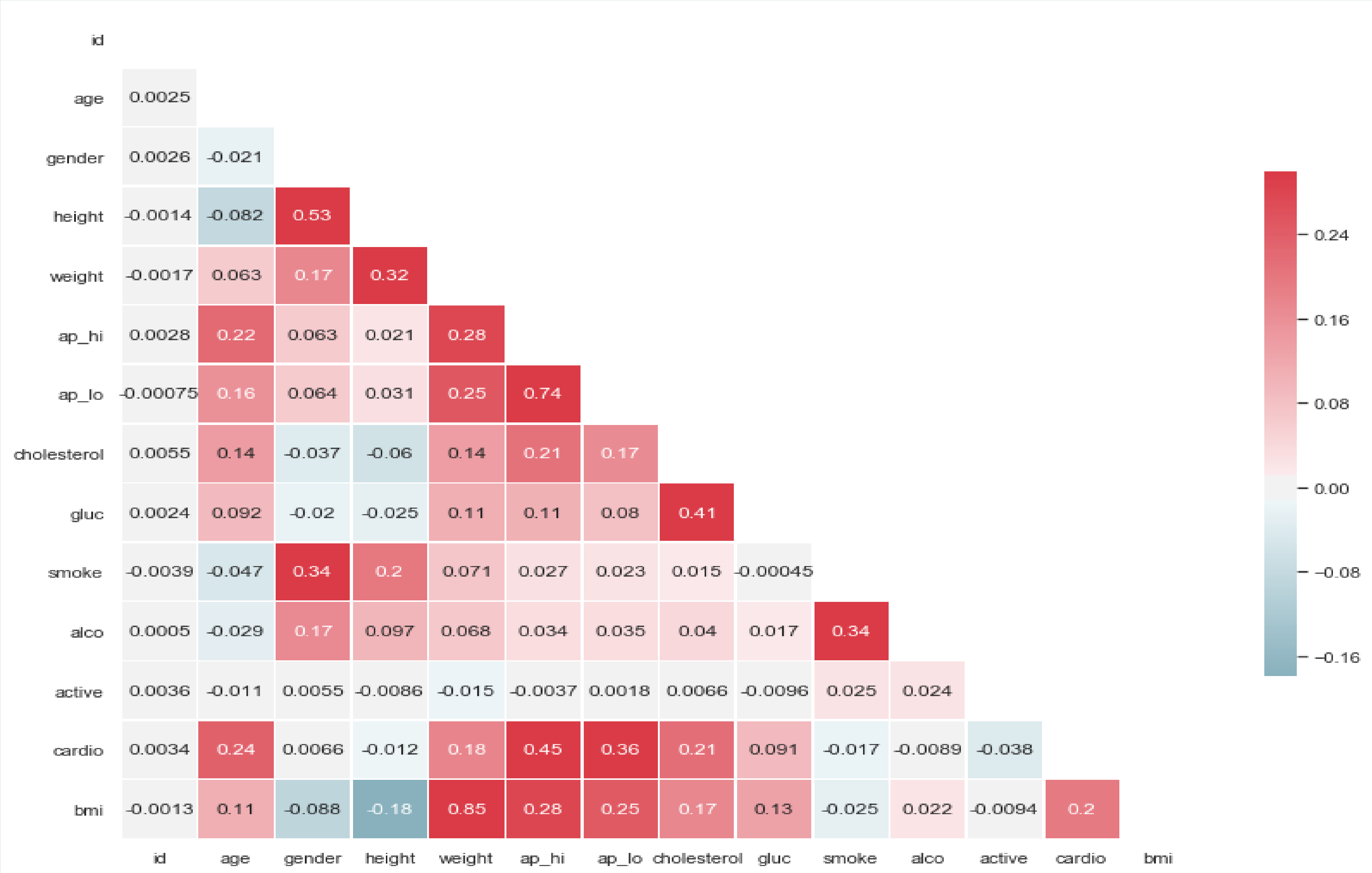
Correlation appears slightly noticeable.

DIASTOLIC - CARDIO

Indicates some positive correlation.

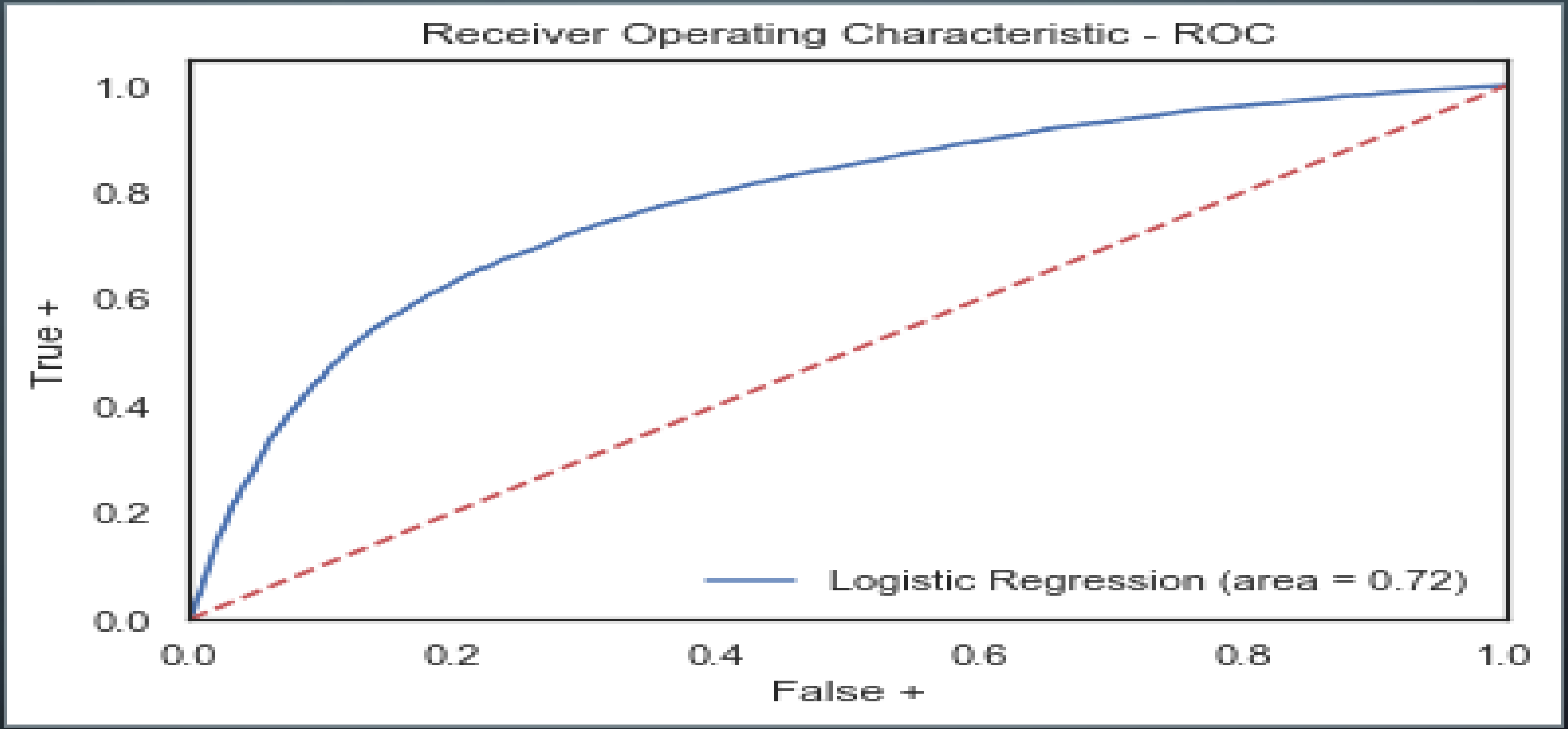
SYSTOLIC - CARDIO

Strongest visual of positive correlation.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi
id														
age	0.0025													
gender	0.0026	-0.021												
height	-0.0014	-0.082	0.53											
weight	-0.0017	0.063	0.17	0.32										
ap_hi	0.0028	0.22	0.063	0.021	0.28									
ap_lo	-0.00075	0.16	0.064	0.031	0.25	0.74								
cholesterol	0.0055	0.14	-0.037	-0.06	0.14	0.21	0.17							
gluc	0.0024	0.092	-0.02	-0.025	0.11	0.11	0.08	0.41						
smoke	-0.0039	-0.047	0.34	0.2	0.071	0.027	0.023	0.015	-0.00045					
alco	0.0005	-0.029	0.17	0.097	0.068	0.034	0.035	0.04	0.017	0.34				
active	0.0036	-0.011	0.0055	-0.0086	-0.015	-0.0037	0.0018	0.0066	-0.0096	0.025	0.024			
cardio	0.0034	0.24	0.0066	-0.012	0.18	0.45	0.36	0.21	0.091	-0.017	-0.0089	-0.038		
bmi	-0.0013	0.11	-0.088	-0.18	0.85	0.28	0.25	0.17	0.13	-0.025	0.022	-0.0094	0.2	

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.