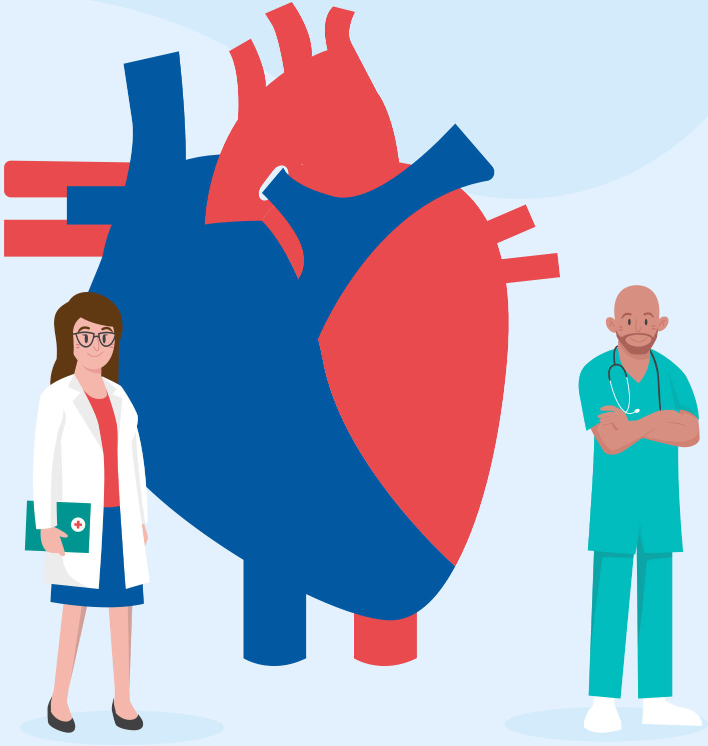
# CARDIOVASCULAR DISEASE

Group 5 - Azfar, Pei Yu, Shi Wen, Zheng Nan



# Presentation Outline

1. Problem statement/Goal
2. Background
3. Data Features/Preparation
4. Data Visualisation
5. Machine Learning
6. Conclusion

# CARDIOVASCULAR DISEASE (CVD)

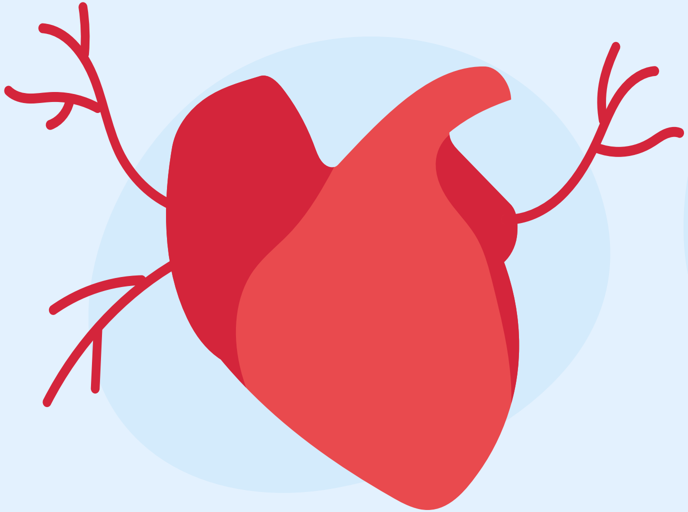
a disease of the heart or blood vessels

1 out of 3 deaths in Singapore is due to cardiovascular diseases

17.9 million people die from CVDs, which estimates to 32% worldwide

# PROBLEM STATEMENT

What are the daily habits that might contribute to the chances of getting cardiovascular disease.



# Typical Stereotypes Habits of CVD

### Smoking

Smoke directly goes to your lungs

### Alcohol

If taken excessively, can cause health complications

### Activeness

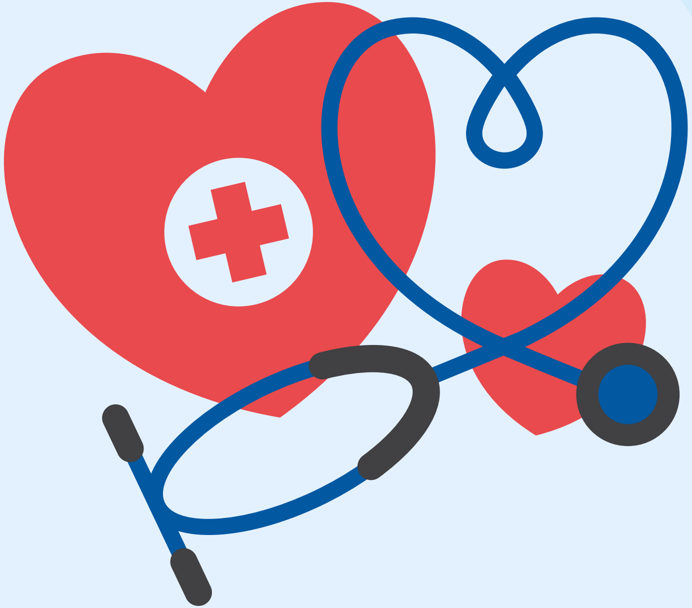
Inactive  = Weak Lungs = Prone to CVD

### Diet

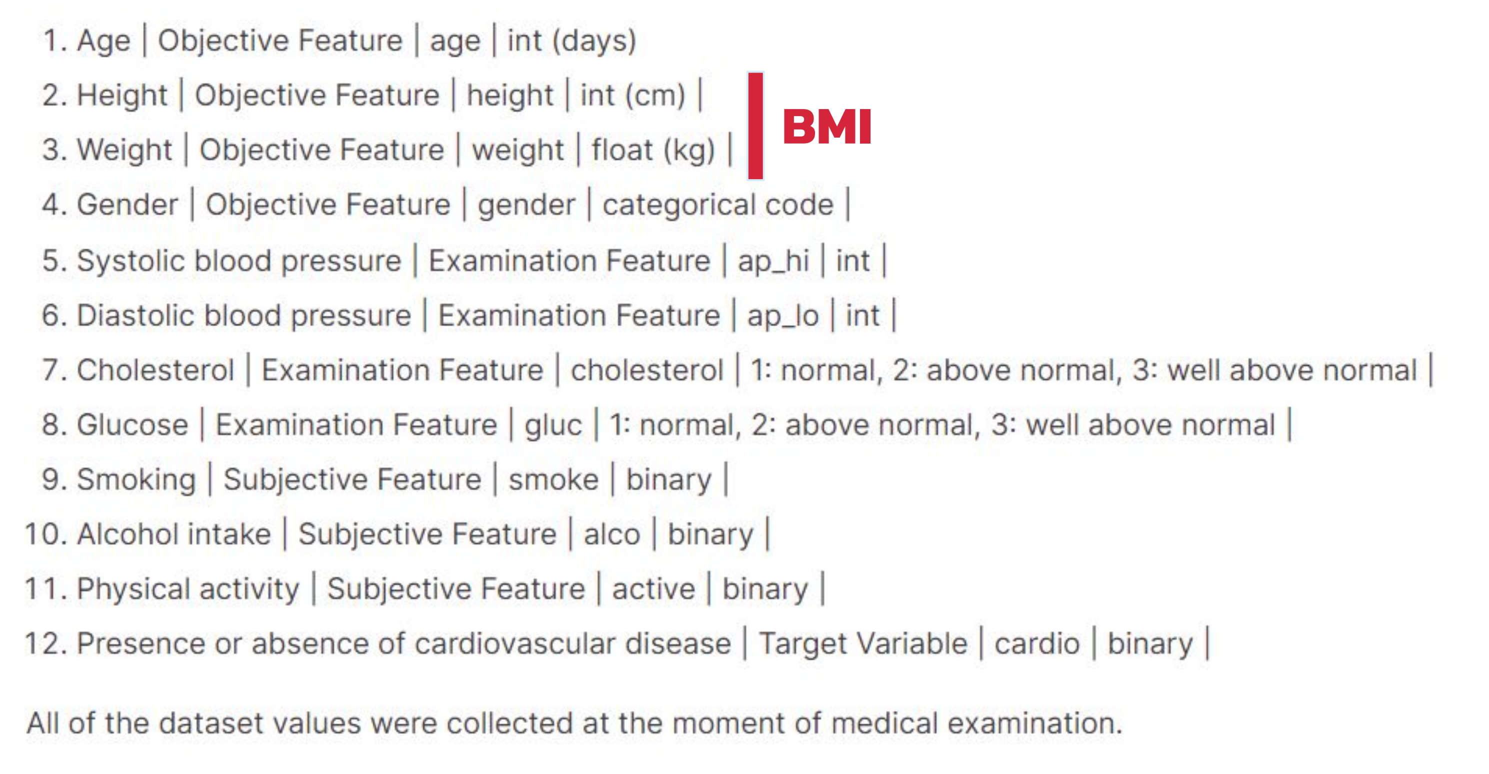
Fatty Diet can cause complications

# OUR GOAL

1. Find out the daily habits that contribute to the chances of getting cardiovascular diseases
2. Solutions to reduce the death rate from cardiovascular diseases by reducing the chance of Cardiovascular disease.



# Data Features

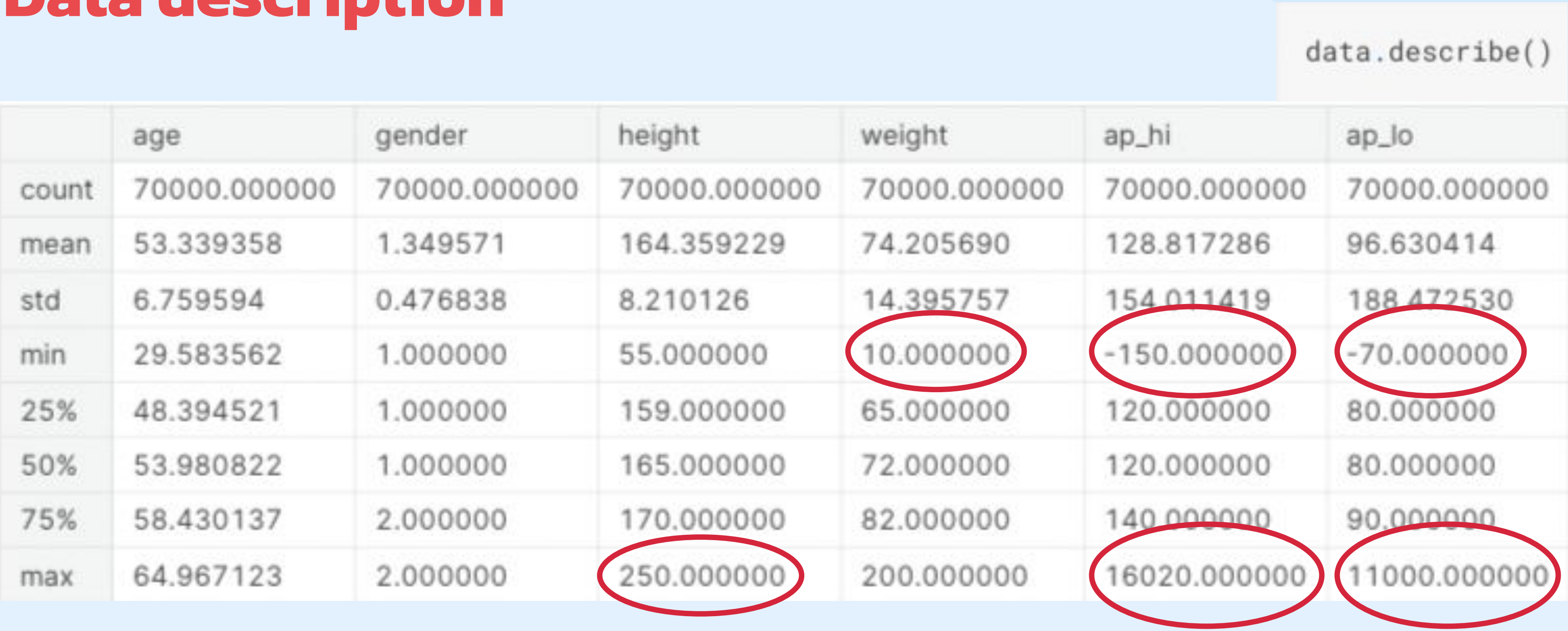


1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

BMI

All of the dataset values were collected at the moment of medical examination.

# Data description



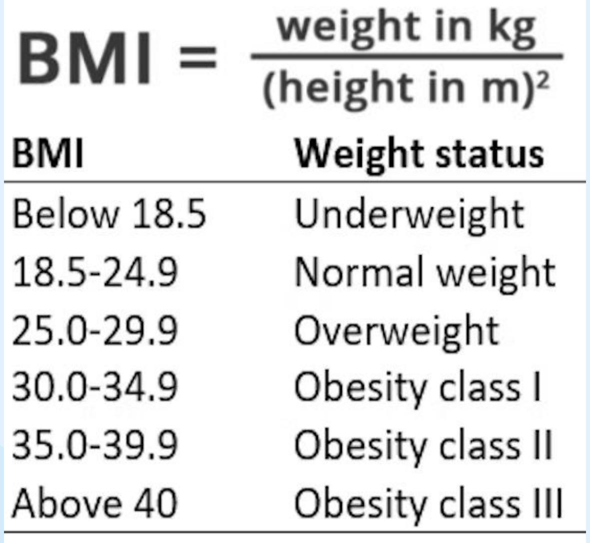
```
data.describe()
```

| | age | gender | height | weight | ap_hi | ap_lo |
|---|---|---|---|---|---|---|
| count | 70000.000000 | 70000.000000 | 70000.000000 | 70000.000000 | 70000.000000 | 70000.000000 |
| mean | 53.339358 | 1.349571 | 164.359229 | 74.205690 | 128.817286 | 96.630414 |
| std | 6.759594 | 0.476838 | 8.210126 | 14.395757 | 154.011419 | 188.472530 |
| min | 29.583562 | 1.000000 | 55.000000 | 10.000000 | -150.000000 | -70.000000 |
| 25% | 48.394521 | 1.000000 | 159.000000 | 65.000000 | 120.000000 | 80.000000 |
| 50% | 53.980822 | 1.000000 | 165.000000 | 72.000000 | 120.000000 | 80.000000 |
| 75% | 58.430137 | 2.000000 | 170.000000 | 82.000000 | 140.000000 | 90.000000 |
| max | 64.967123 | 2.000000 | 250.000000 | 200.000000 | 16020.000000 | 11000.000000 |

**Can be seen that there are unreasonable blood pressure values (negative values) as well as height and weight values.**

# Basic information

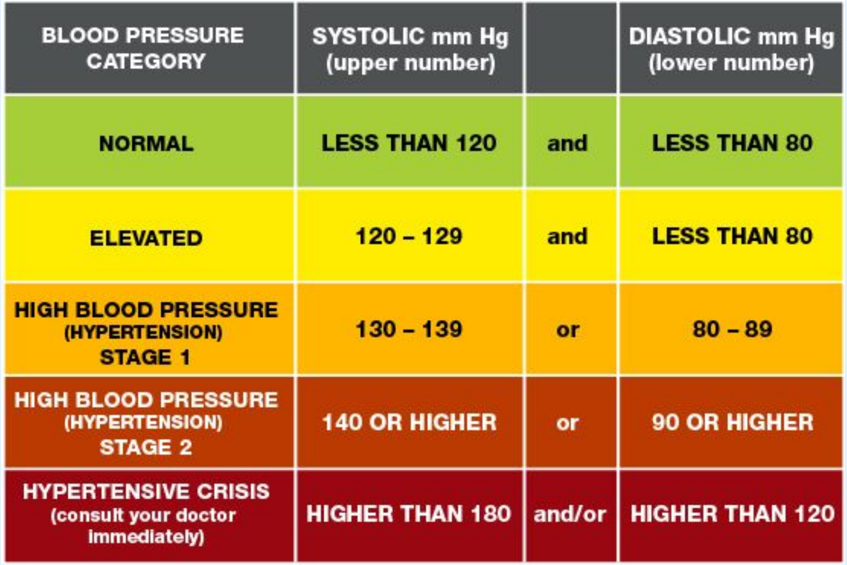
## BMI Ranges



$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in m})^2}$$

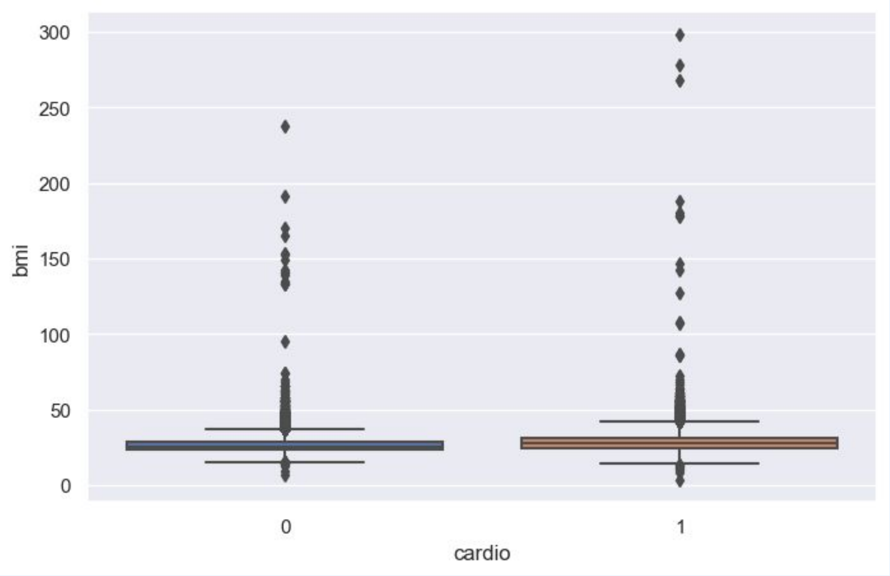
| BMI | Weight status |
|---|---|
| Below 18.5 | Underweight |
| 18.5-24.9 | Normal weight |
| 25.0-29.9 | Overweight |
| 30.0-34.9 | Obesity class I |
| 35.0-39.9 | Obesity class II |
| Above 40 | Obesity class III |

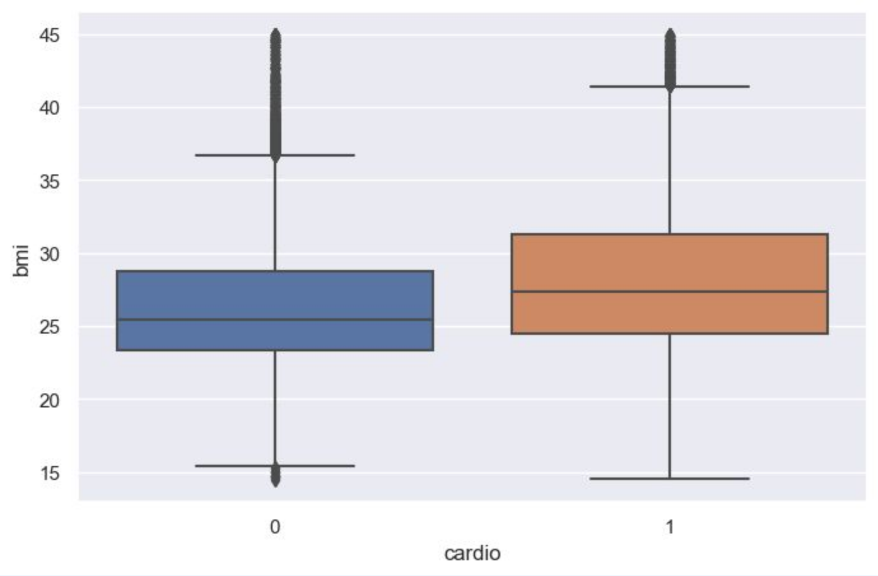
## Blood Pressure Range



| BLOOD PRESSURE CATEGORY | SYSTOLIC mm Hg (upper number) | | DIASTOLIC mm Hg (lower number) |
|---|---|---|---|
| NORMAL | LESS THAN 120 | and | LESS THAN 80 |
| ELEVATED | 120 – 129 | and | LESS THAN 80 |
| HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1 | 130 – 139 | or | 80 – 89 |
| HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2 | 140 OR HIGHER | or | 90 OR HIGHER |
| HYPERTENSIVE CRISIS (consult your doctor immediately) | HIGHER THAN 180 | and/or | HIGHER THAN 120 |

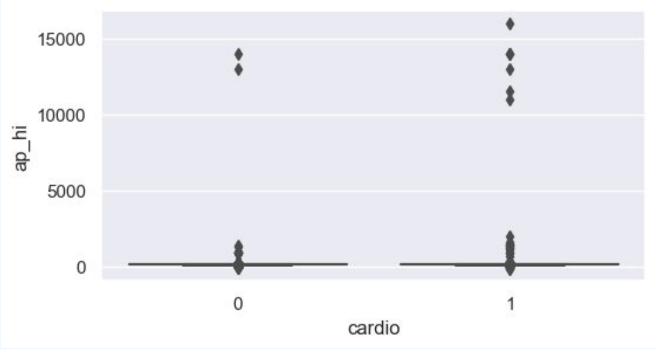
# Data comparison (BMI)

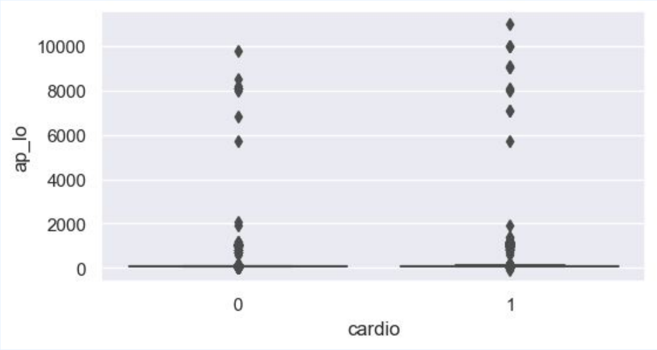




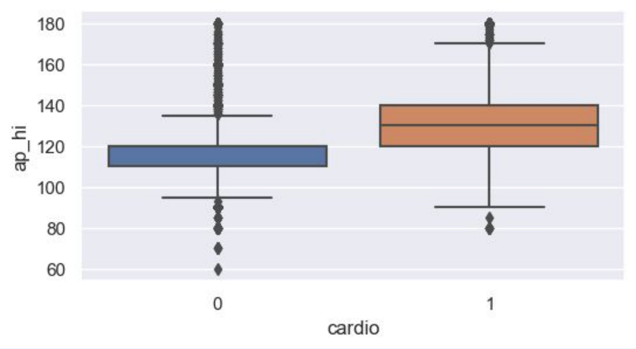
# Data comparison (Blood Pressure)

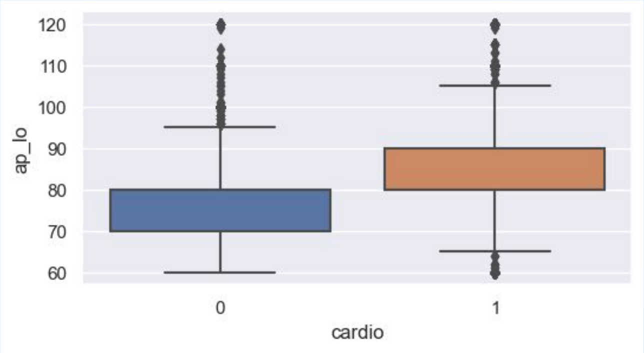
## Before



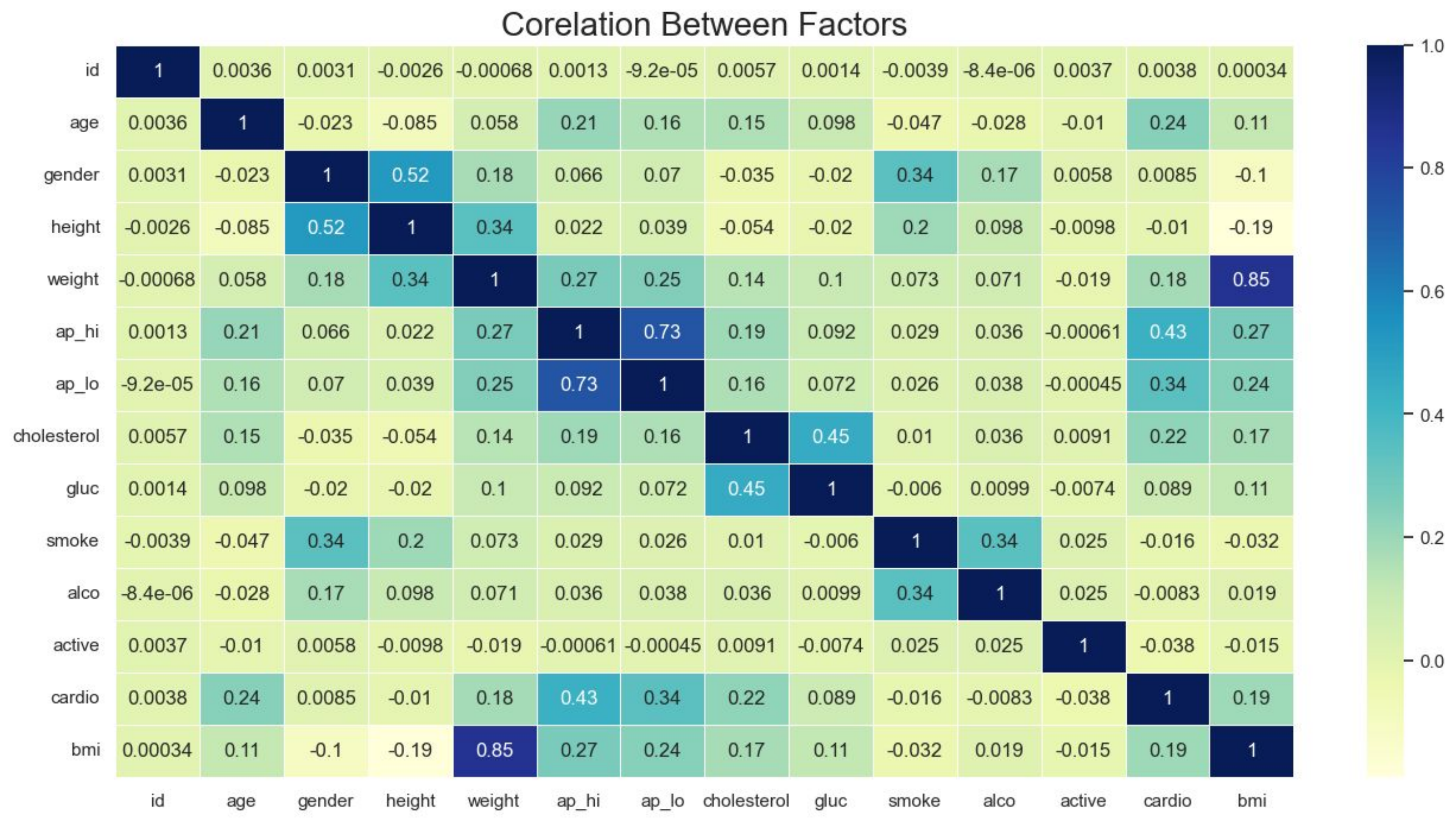


## After





# Correlation Heatmap



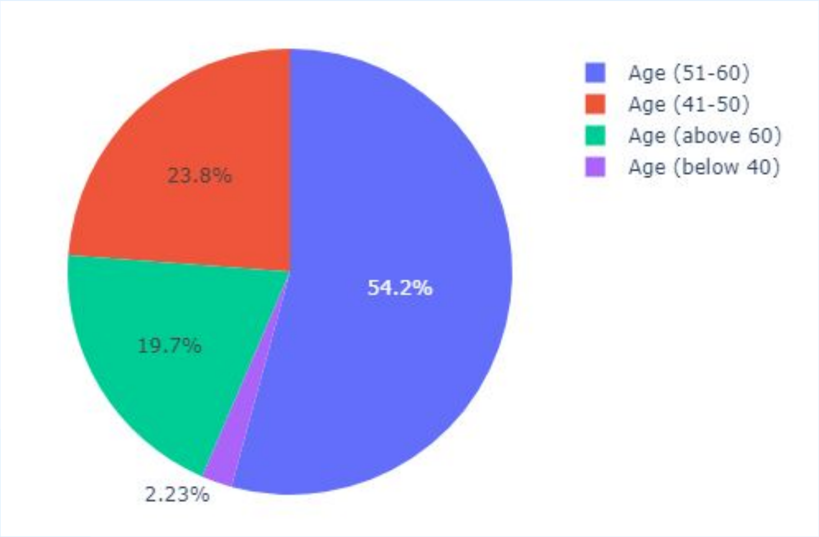
Corelation Between Factors

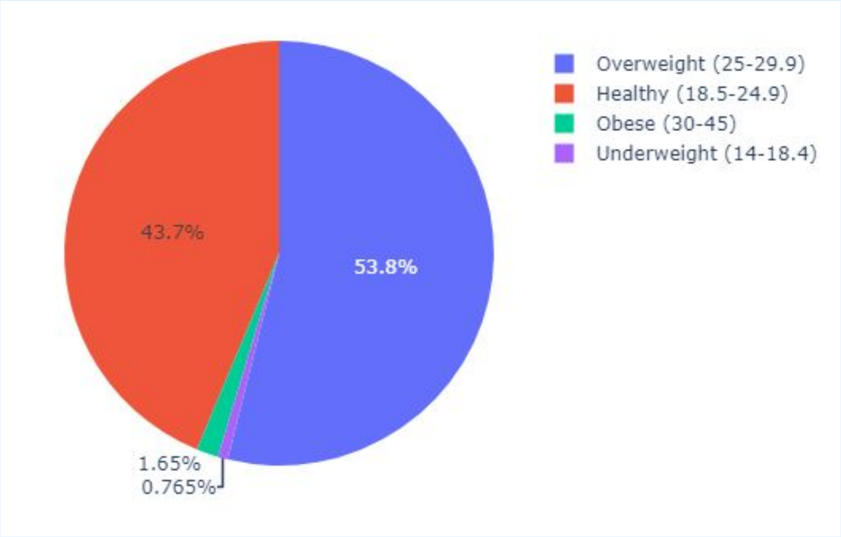
| | id | age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio | bmi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| id | 1 | 0.0036 | 0.0031 | -0.0026 | -0.00068 | 0.0013 | -9.2e-05 | 0.0057 | 0.0014 | -0.0039 | -8.4e-06 | 0.0037 | 0.0038 | 0.00034 |
| age | 0.0036 | 1 | -0.023 | -0.085 | 0.058 | 0.21 | 0.16 | 0.15 | 0.098 | -0.047 | -0.028 | -0.01 | 0.24 | 0.11 |
| gender | 0.0031 | -0.023 | 1 | 0.52 | 0.18 | 0.066 | 0.07 | -0.035 | -0.02 | 0.34 | 0.17 | 0.0058 | 0.0085 | -0.1 |
| height | -0.0026 | -0.085 | 0.52 | 1 | 0.34 | 0.022 | 0.039 | -0.054 | -0.02 | 0.2 | 0.098 | -0.0098 | -0.01 | -0.19 |
| weight | -0.00068 | 0.058 | 0.18 | 0.34 | 1 | 0.27 | 0.25 | 0.14 | 0.1 | 0.073 | 0.071 | -0.019 | 0.18 | 0.85 |
| ap_hi | 0.0013 | 0.21 | 0.066 | 0.022 | 0.27 | 1 | 0.73 | 0.19 | 0.092 | 0.029 | 0.036 | -0.00061 | 0.43 | 0.27 |
| ap_lo | -9.2e-05 | 0.16 | 0.07 | 0.039 | 0.25 | 0.73 | 1 | 0.16 | 0.072 | 0.026 | 0.038 | -0.00045 | 0.34 | 0.24 |
| cholesterol | 0.0057 | 0.15 | -0.035 | -0.054 | 0.14 | 0.19 | 0.16 | 1 | 0.45 | 0.01 | 0.036 | 0.0091 | 0.22 | 0.17 |
| gluc | 0.0014 | 0.098 | -0.02 | -0.02 | 0.1 | 0.092 | 0.072 | 0.45 | 1 | -0.006 | 0.0099 | -0.0074 | 0.089 | 0.11 |
| smoke | -0.0039 | -0.047 | 0.34 | 0.2 | 0.073 | 0.029 | 0.026 | 0.01 | -0.006 | 1 | 0.34 | 0.025 | -0.016 | -0.032 |
| alco | -8.4e-06 | -0.028 | 0.17 | 0.098 | 0.071 | 0.036 | 0.038 | 0.036 | 0.0099 | 0.34 | 1 | 0.025 | -0.0083 | 0.019 |
| active | 0.0037 | -0.01 | 0.0058 | -0.0098 | -0.019 | -0.00061 | -0.00045 | 0.0091 | -0.0074 | 0.025 | 0.025 | 1 | -0.038 | -0.015 |
| cardio | 0.0038 | 0.24 | 0.0085 | -0.01 | 0.18 | 0.43 | 0.34 | 0.22 | 0.089 | -0.016 | -0.0083 | -0.038 | 1 | 0.19 |
| bmi | 0.00034 | 0.11 | -0.1 | -0.19 | 0.85 | 0.27 | 0.24 | 0.17 | 0.11 | -0.032 | 0.019 | -0.015 | 0.19 | 1 |

# Main Features with high correlation against CVD

| | Features | Correlation |
|---|---|---|
| 1 | ap_hi (Systolic blood pressure) | 0.43 |
| 2 | ap_lo (Diastolic blood pressure) | 0.34 |
| 3 | Cholesterol | 0.22 |
| 4 | Age | 0.24 |
| 5 | BMI | 0.19 |

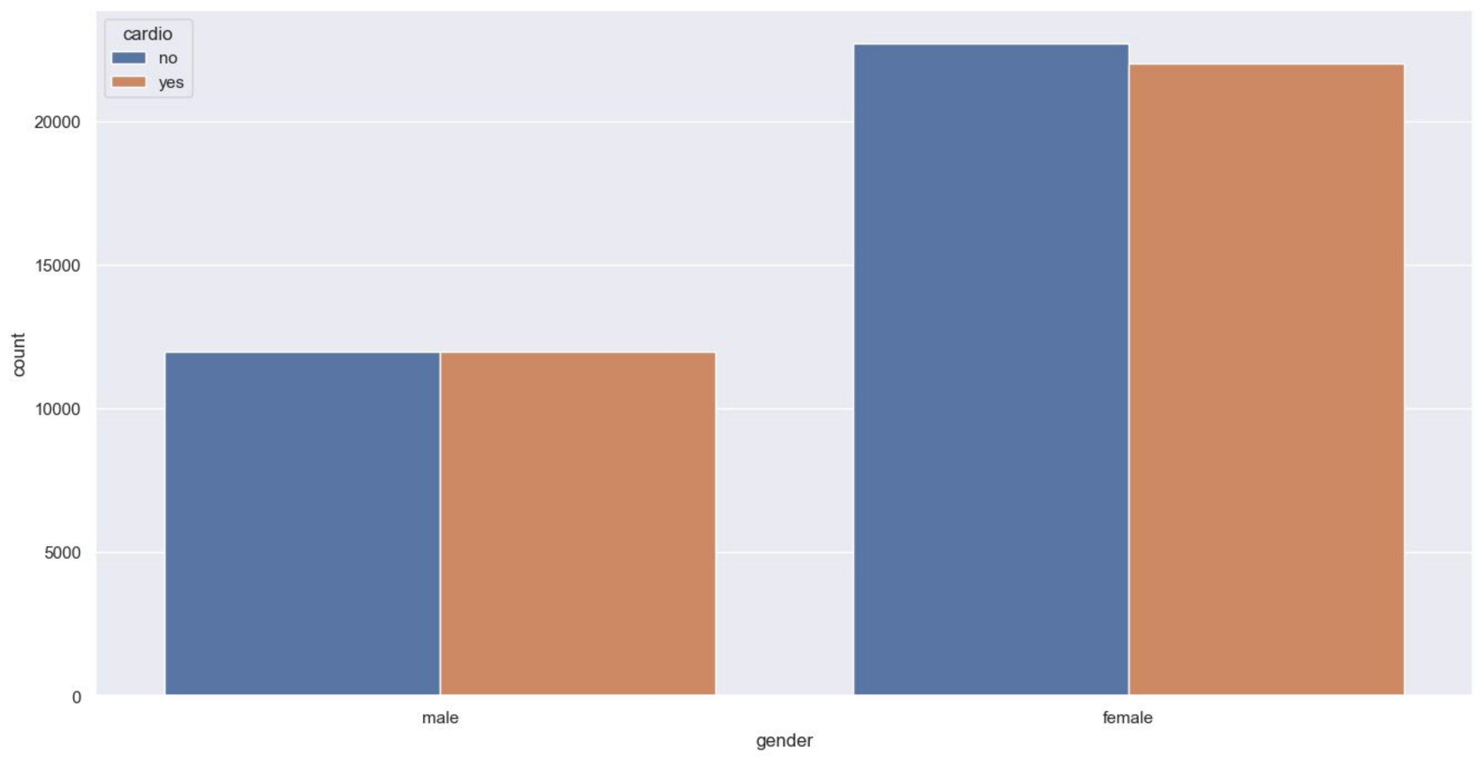
# Cardio vs Age & BMI



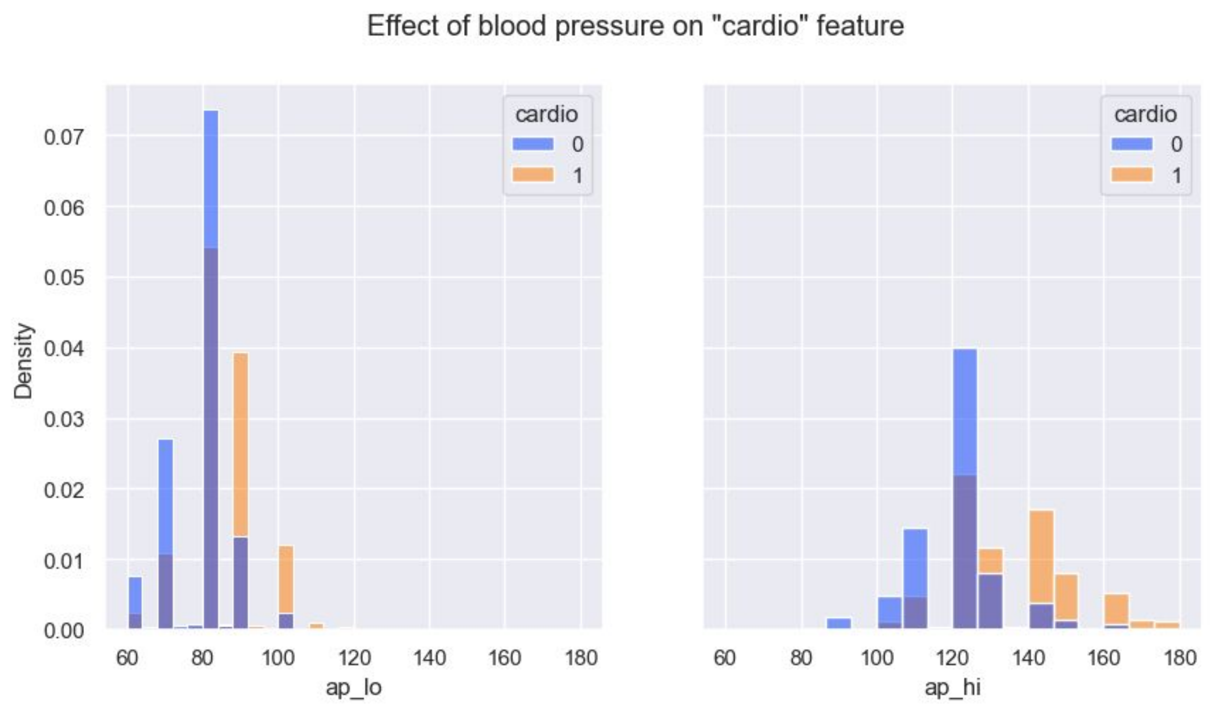


**Pie chart is produced using Plotly**

# Cardio vs Gender

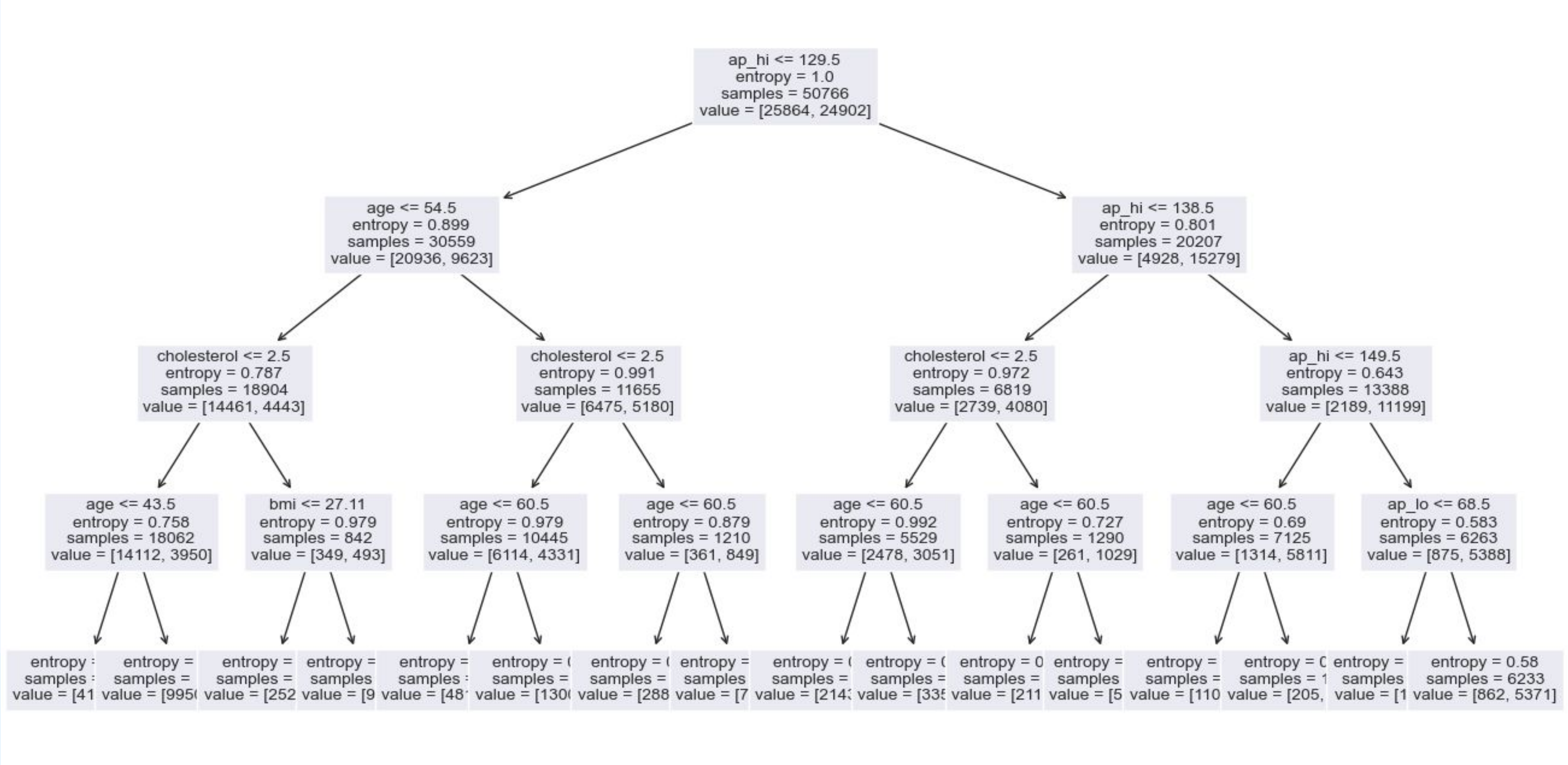


# Cardio vs Blood Pressure



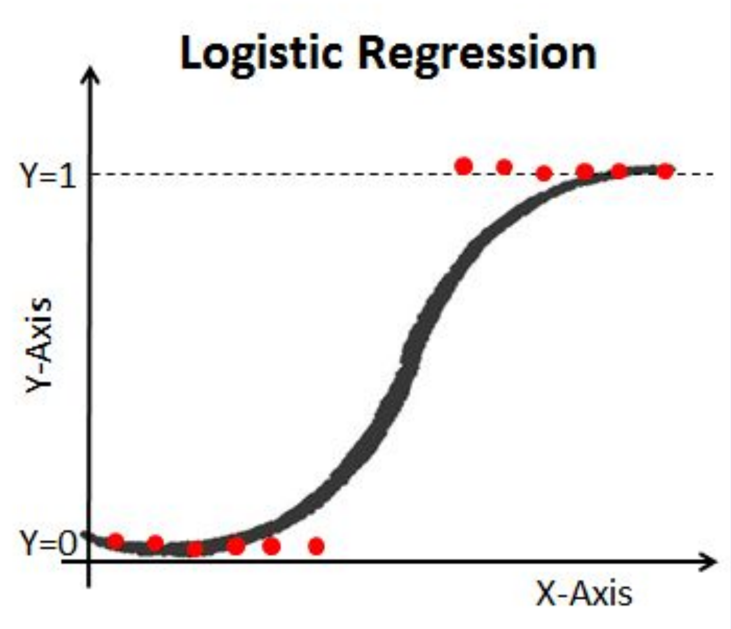
Higher ap_hi and ap_lo results in higher incidence of CVDs

# Decision Tree

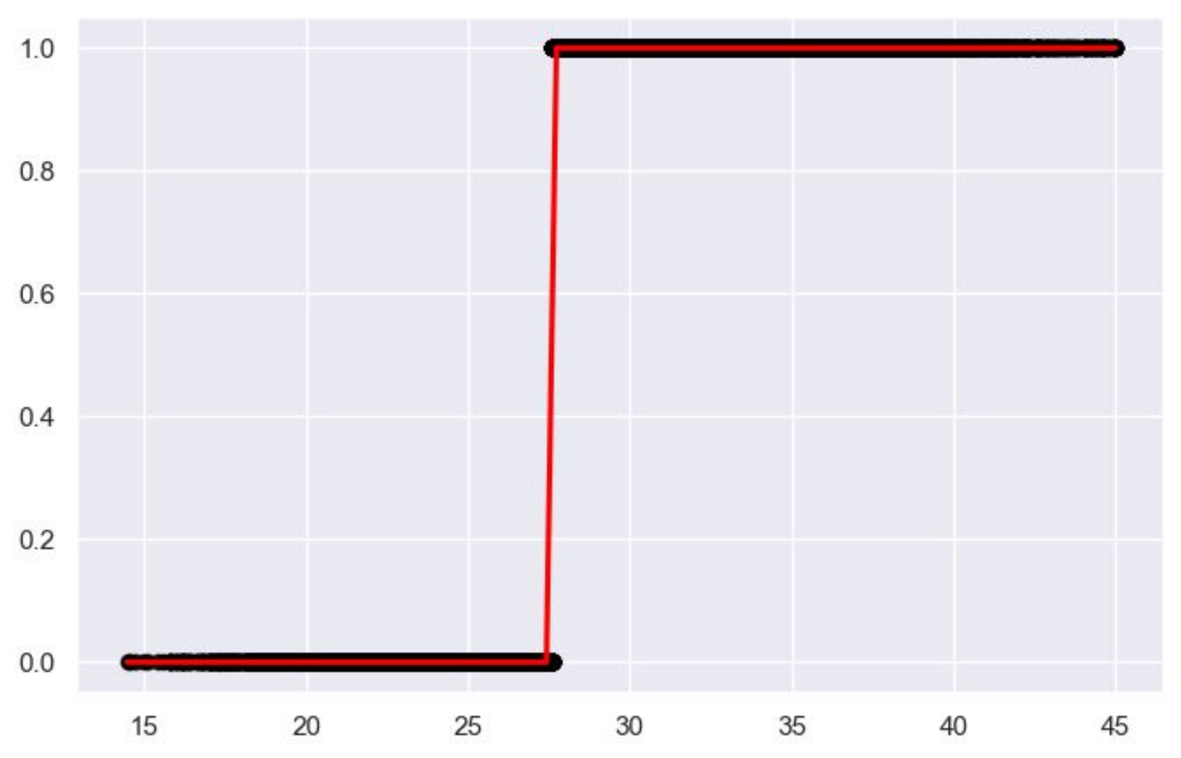


# Logistic Regression

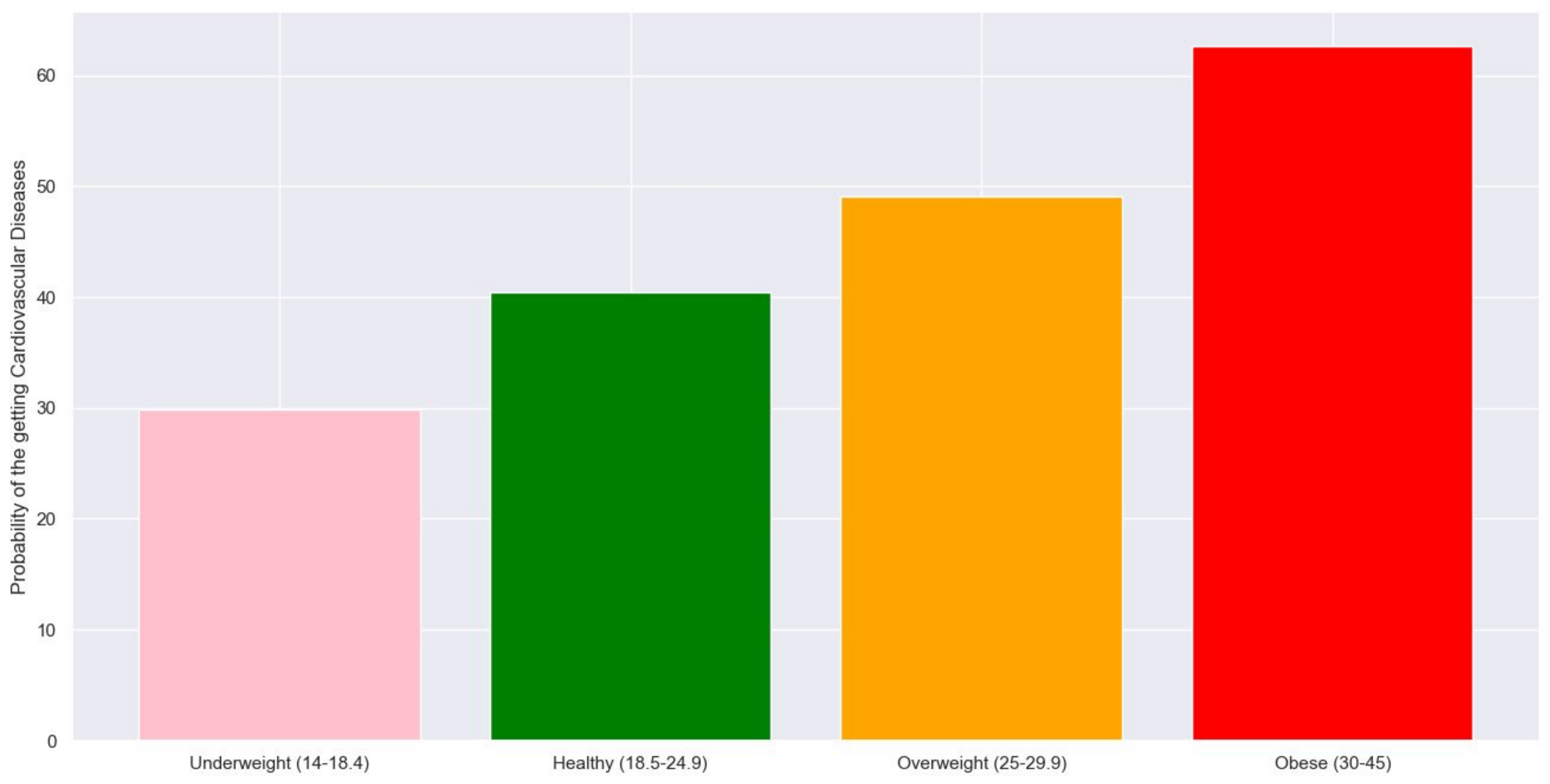
- A logit function is used to predict the likelihood of occurrence of a binary event
- Estimates the relationship between one dependent binary variable and independent variables.
- Estimation is done by Maximum Likelihood



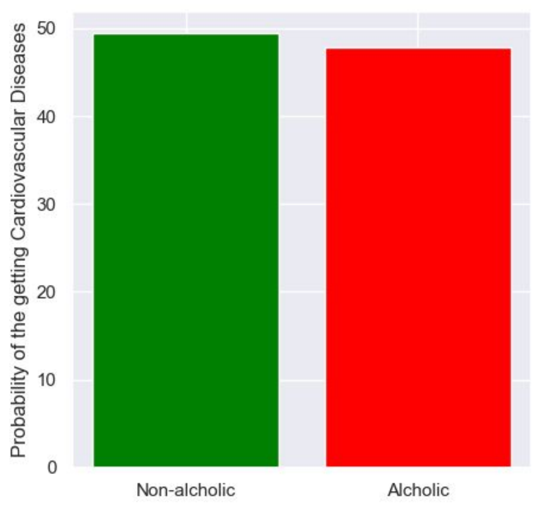
# Logistic Regression Graph - BMI vs Cardio

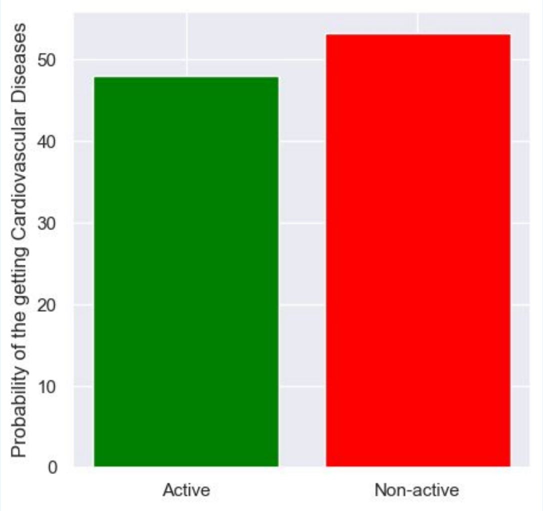


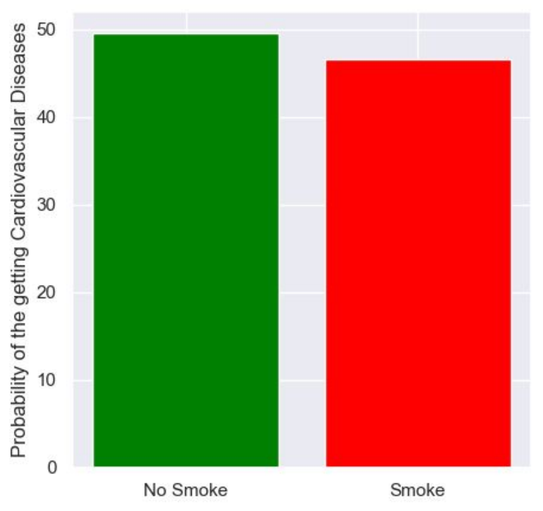
# Probability of Getting Cardio (BMI) - Logistic Regression



# Probability of Getting Cardio – Logistic Regression

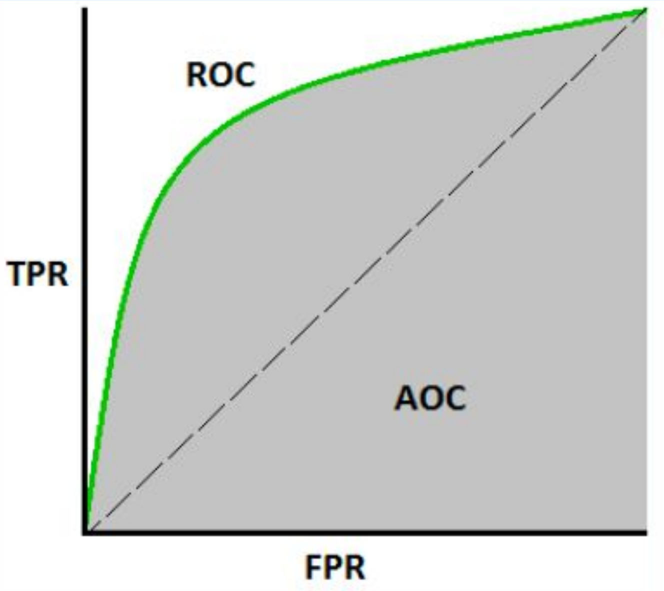




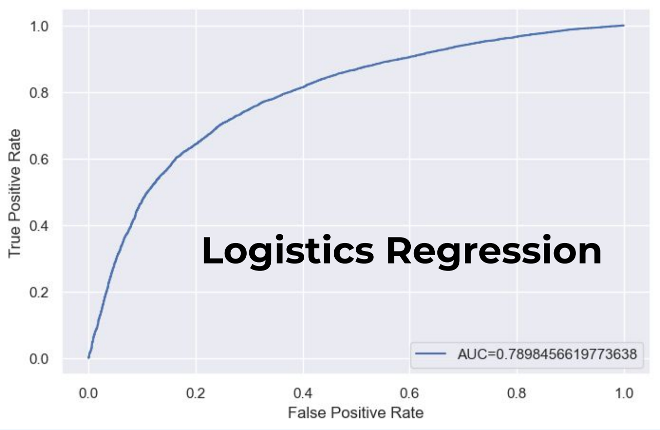


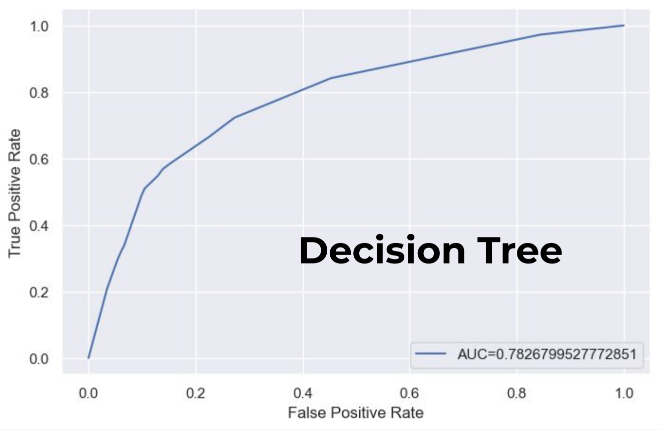
# ROC Curve

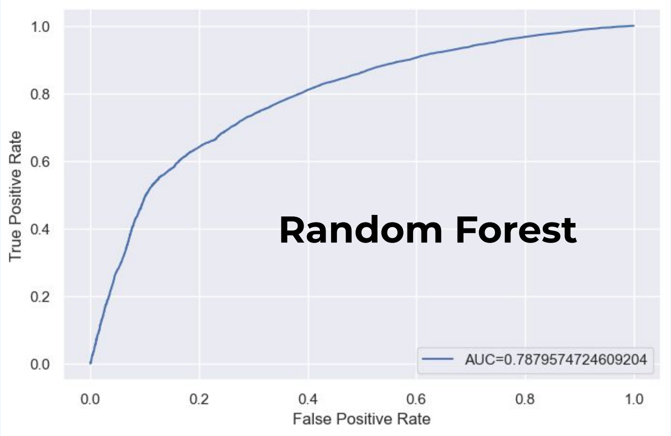
- Shows the performance of machine learning model
- The sharper the angle of the curve the better the model
- AUC determines the performance of the machine learning model



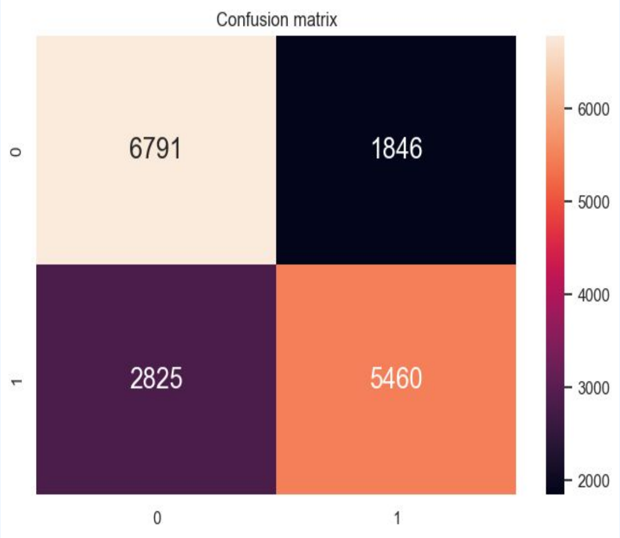
# ROC Curve with AUC

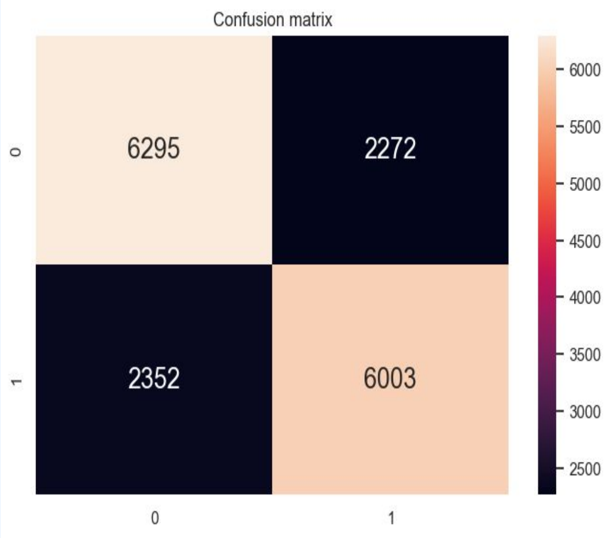

Logistics Regression

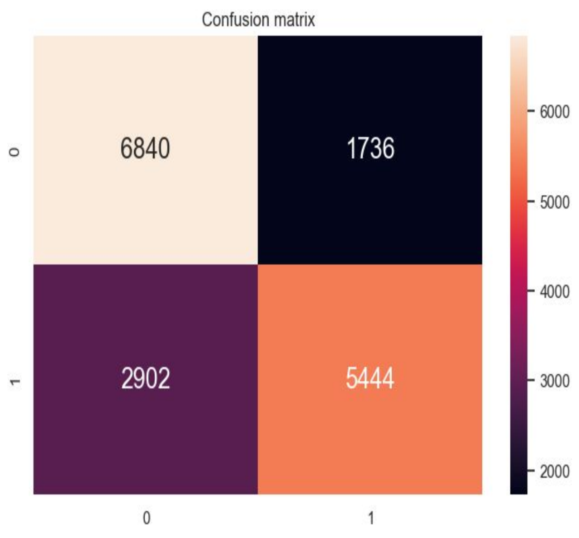

Decision Tree

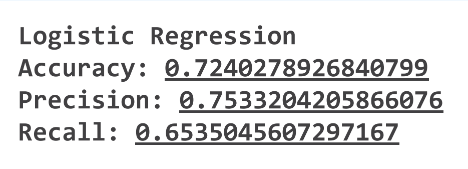

Random Forest

# Confusion Matrix

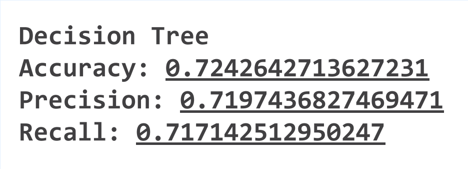




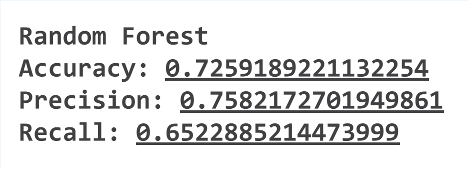




```
Logistic Regression
Accuracy: 0.7240278926840799
Precision: 0.7533204205866076
Recall: 0.6535045607297167
```



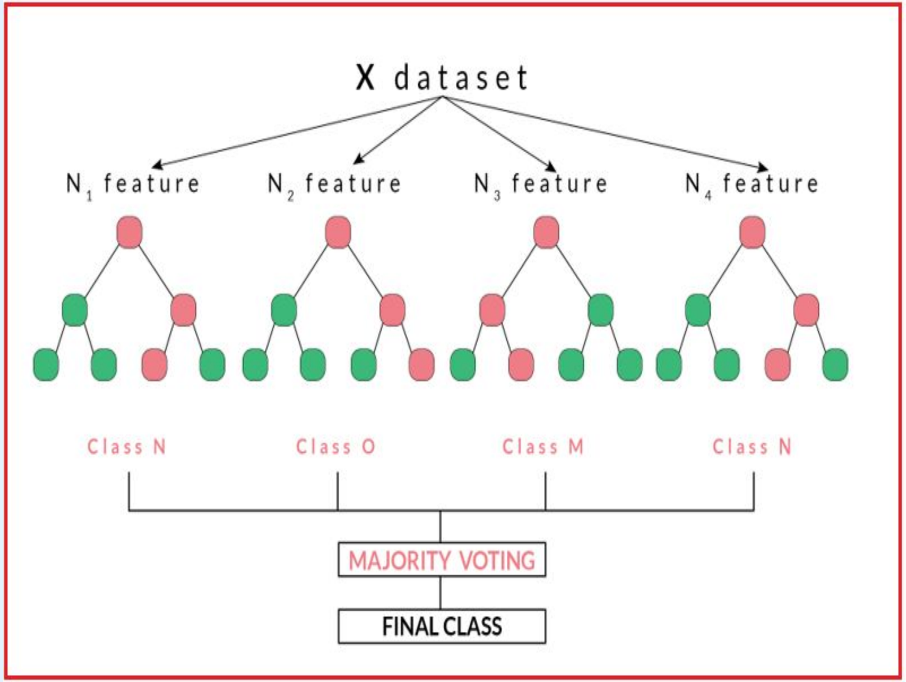
```
Decision Tree
Accuracy: 0.7242642713627231
Precision: 0.7197436827469471
Recall: 0.717142512950247
```



```
Random Forest
Accuracy: 0.7259189221132254
Precision: 0.7582172701949861
Recall: 0.6522885214473999
```

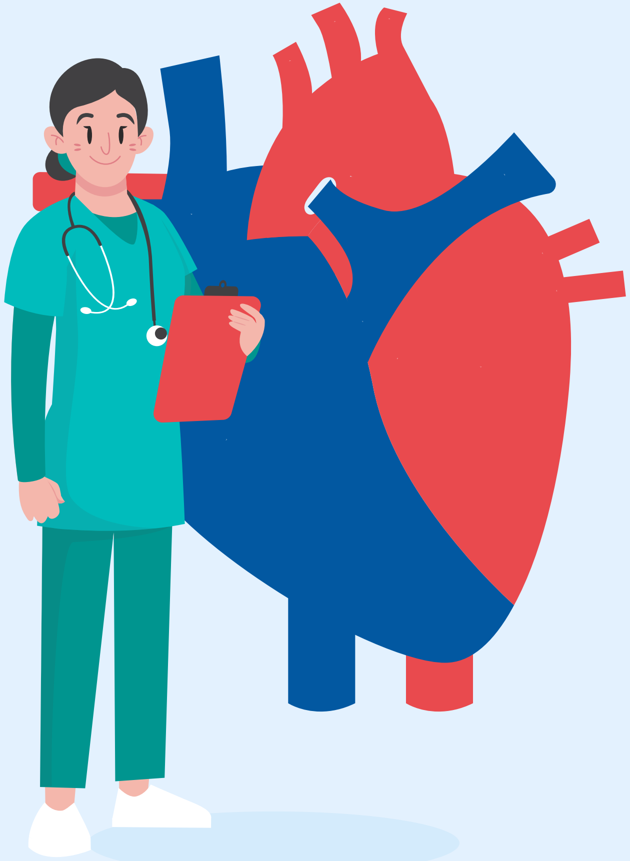
# Random Forest

- Uses multiple decision trees for prediction
- Final result is based on majority voting/Averaging
- Train data is used to create the decision trees
- Can perform both regression and classification tasks
- Handle large dataset efficiently
- Produces good predictions



# Comparison of the models used

| Model | Accuracy |
|---|---|
| Decision Tree | 0.724 |
| Logistic Regression | 0.724 |
| Random Forest | 0.725 |



# Conclusion

- Not all typical stereotype habits have significant impact to CVD
- Blood Pressure and BMI have the strongest relationship with the presence of CVD
- Having a healthy diet would significantly reduce the chances of CVD

# Our Team

**Azfar**

- Decision Tree
- Logistic Regression
- Data Preparation
- Data Exploratory

**Shi Wen**

- Data preparation
- Data Exploratory
- Decision Tree

**Zheng Nan**

- Data Preparation
- Data Exploratory
- RandomForest

**Pei Yu**

- Logistic Regression
- Plotly
- Data Preparation
- Data Exploratory

# Thanks!

# Q&A

# Reference

Shafi, A. (2018, May 16). Random Forest classification with Scikit-Learn. DataCamp. Retrieved April 13, 2023, from https://www.datacamp.com/tutorial/random-forests-classifier-python

Navlani, A. (2018, December 28). Python decision tree classification tutorial: Scikit-Learn Decisiontreeclassifier. DataCamp. Retrieved April 13, 2023, from https://www.datacamp.com/tutorial/decision-tree-classification-python

Ulianova, S. (2019, January 20). Cardiovascular disease dataset. Kaggle. Retrieved April 13, 2023, from https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Navlani, A. (2019, December 16). Python logistic regression tutorial with Sklearn &amp; Scikit. DataCamp. Retrieved April 13, 2023, from https://www.datacamp.com/tutorial/understanding-logistic-regression-python?utm_source=google&amp;utm_medium=paid_search&amp;utm_campaignid=19589720821&amp;utm_adgroupid=143216588577&amp;utm_device=c&amp;utm_keyword=&amp;utm_matchtype=&amp;utm_network=g&amp;utm_adpostion=&amp;utm_creative=652967469592&amp;utm_targetid=aud-299261629574%3Adsa-1947282172981&amp;utm_loc_interest_ms=&amp;utm_loc_physical_ms=9062499&amp;utm_content=dsa~page~community-tuto&amp;utm_campaign=230119_1-sea~dsa~tutorials_2-b2c_3-row-p1_4-prc_5-na_6-na_7-le_8-pdsh-go_9-na_10-na_11-na&amp;gclid=CjwKCAjw0N6hBhAUEiwAXab-TXFJKgP_JbRA1JRfup3lfRQGTzUa7z95YUnPynmE72joHXr-7N3o9BoCJNIQAvD_BwE

Pie. Pie charts in Python. (n.d.). Retrieved April 13, 2023, from https://plotly.com/python/pie-charts/