CIS 2423
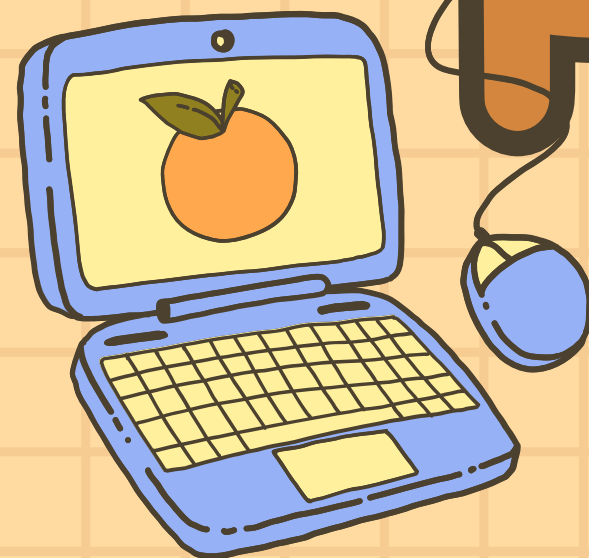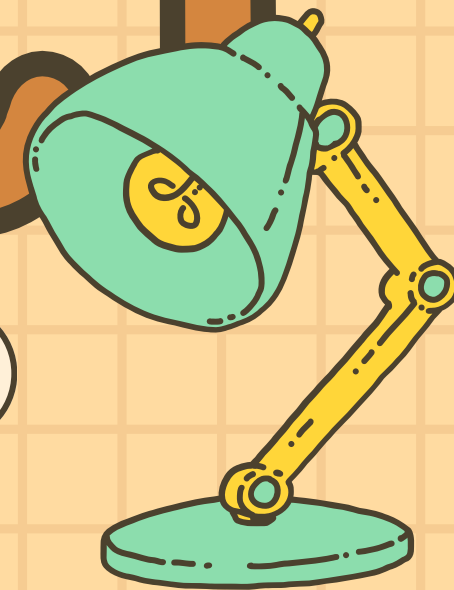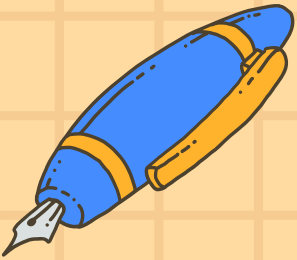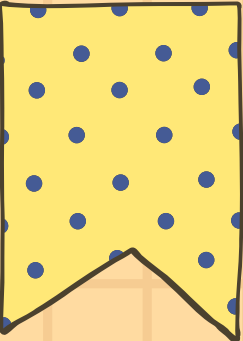
PROJECT

Asma H00492811, HIND H00532324, THAMNA H00491950

INTRODUCTION
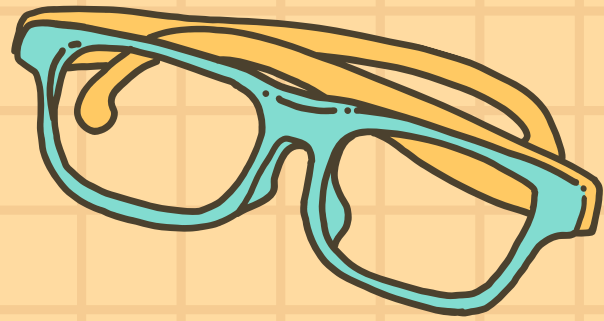
This project studies a dataset with body fat percentages and body measurements. The goal is to predict body fat percentage using python and analyse patterns created from visualisations and clustering in the data. This analysis is done using Python and includes regression, classification, and clustering models.

Language: Python.
Libraries: listed at the last page.
Version Control: GitHub.
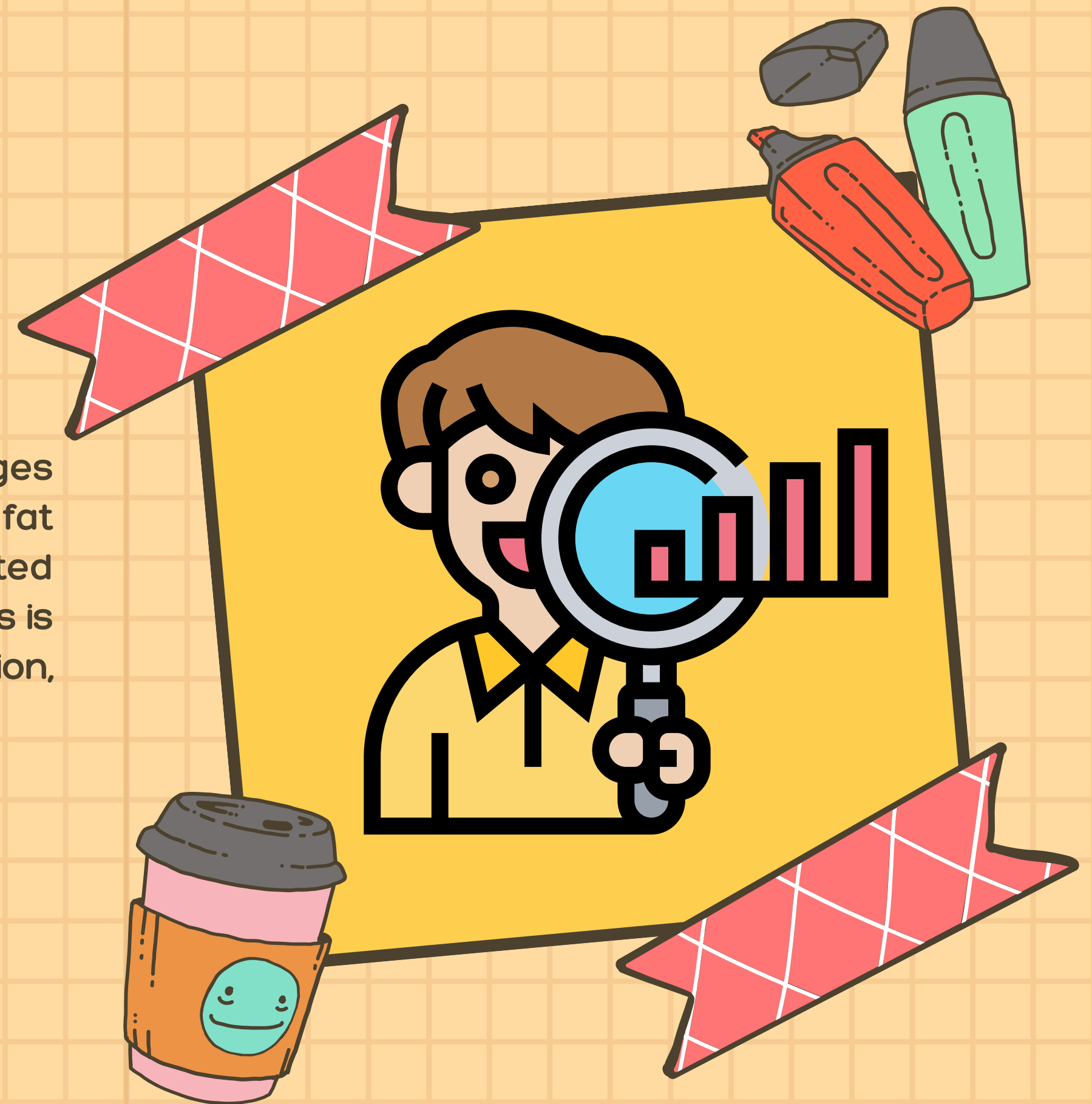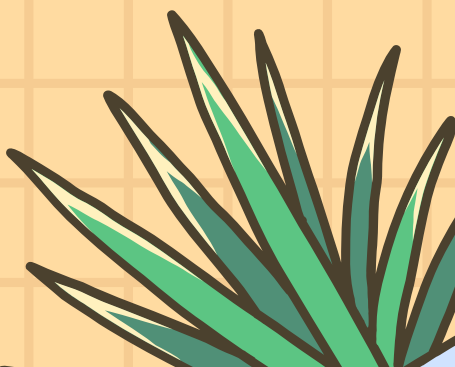
DATA

DATASET: BODYFAT.CSV

-RECORDS: 252.

-VARIABLES: 15 (DENSITY, BODYFAT, AGE, WEIGHT, HEIGHT, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, FOREARM, WRIST).

-DEPENDENT VARIABLE: BODYFAT.

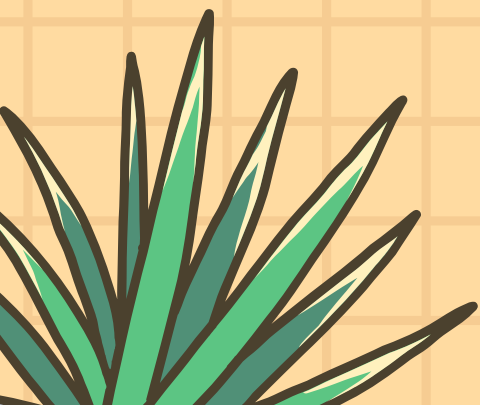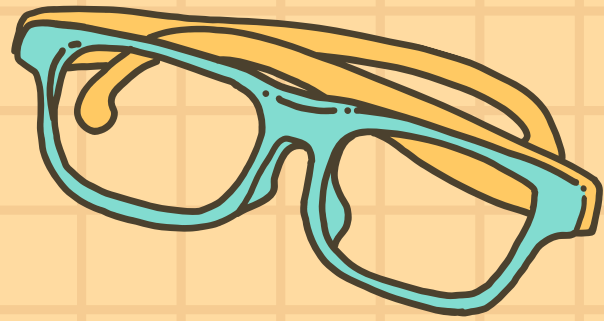-INDEPENDENT VARIABLES: (DENSITY, AGE, WEIGHT, HEIGHT, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, FOREARM, WRIST).

-PURPOSE: PREDICT BODY FAT FROM BODY MEASUREMENTS.

-CHOSEN MACHINE LEARNING ALGORITHMS: LINEAR REGRESSION, LOGISTIC REGRESSION, KNN, NAIVE BAYES, DECISION TREE.
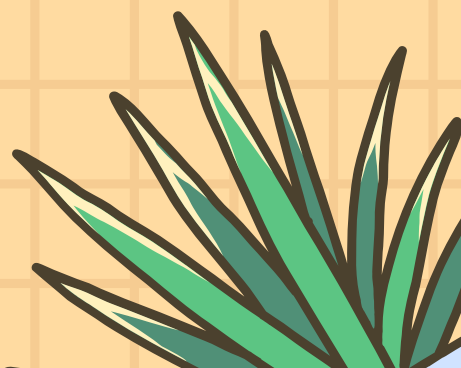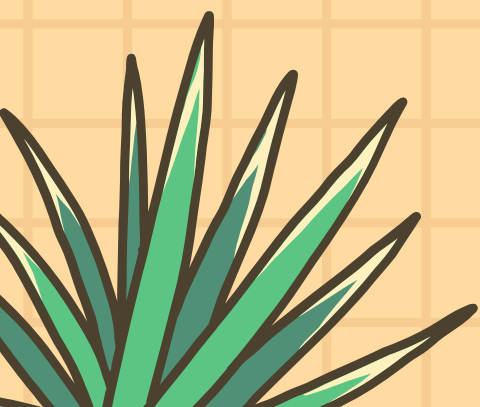
TOP CORRELATIONS WITH BODYFAT

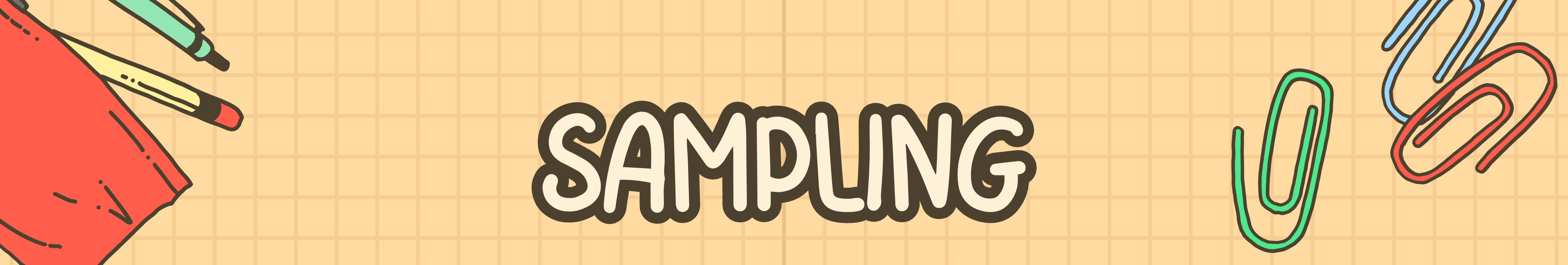| ABDOMEN	CHEST	HIP	WEIGHT
81%	70%	63%	61%

DESCRIPTIVE ANALYTICS

The count suggests complete data. The mean of the bodyfat which is 19% suggests the people in this dataset have a lower than average bodyfat percentage. There seems to be an error in the dataset itself because it states that the minimum bodyfat is 0 and this is impossible. The max bodyfat is generally considered healthy.

|	count	mean	std	min	25%	50%	75%	max
Density	252.0	1.055574	0.019031	0.995	1.0414	1.0549	1.0704	1.1089
BodyFat	252.0	19.150794	8.368740	0.000	12.4750	19.2000	25.3000	47.5000
Age	252.0	44.884921	12.602040	22.000	35.7500	43.0000	54.0000	81.0000
Weight	252.0	178.924405	29.389160	118.500	159.0000	176.5000	197.0000	363.1500
Height	252.0	70.148810	3.662856	29.500	68.2500	70.0000	72.2500	77.7500
Neck	252.0	37.992063	2.430913	31.100	36.4000	38.0000	39.4250	51.2000
Chest	252.0	100.824206	8.430476	79.300	94.3500	99.6500	105.3750	136.2000
Abdomen	252.0	92.555952	10.783077	69.400	84.5750	90.9500	99.3250	148.1000
Hip	252.0	99.904762	7.164058	85.000	95.5000	99.3000	103.5250	147.7000
Thigh	252.0	59.405952	5.249952	47.200	56.0000	59.0000	62.3500	87.3000
Knee	252.0	38.590476	2.411805	33.000	36.9750	38.5000	39.9250	49.1000
Ankle	252.0	23.102381	1.694893	19.100	22.0000	22.8000	24.0000	33.9000
Biceps	252.0	32.273413	3.021274	24.800	30.2000	32.0500	34.3250	45.0000
Forearm	252.0	28.663889	2.020691	21.000	27.3000	28.7000	30.0000	34.9000
Wrist	252.0	18.229762	0.933585	15.800	17.6000	18.3000	18.8000	21.4000

SAMPLING

-Random sampling: 150 records were selected randomly. The average body fat is 19.08, minimum body fat is 0, which is impossibly low, could be an error or missing data. The median is19.9

-Systematic sampling: Selected every kth element (every 1 element) because our dataset is small. This sample includes 150 people with an average body fat of about 19. The values range from 3.7 to 40.1, which is more realistic than the previous sample. Most people have body fat between 13 and 25 percent, with the median at 19.5.
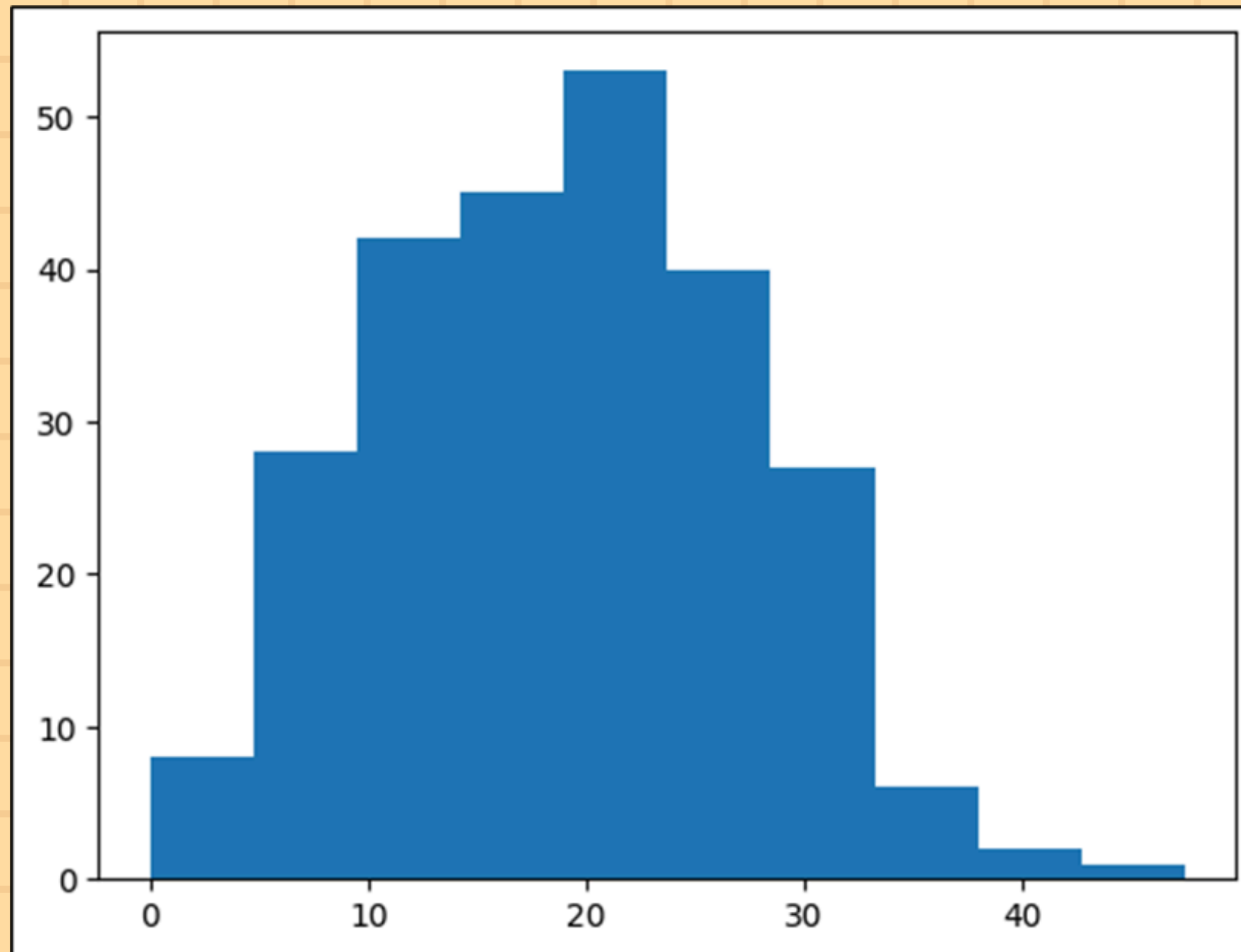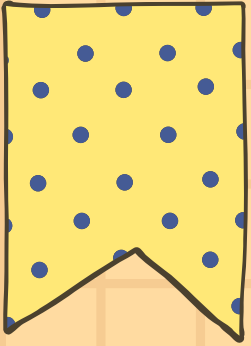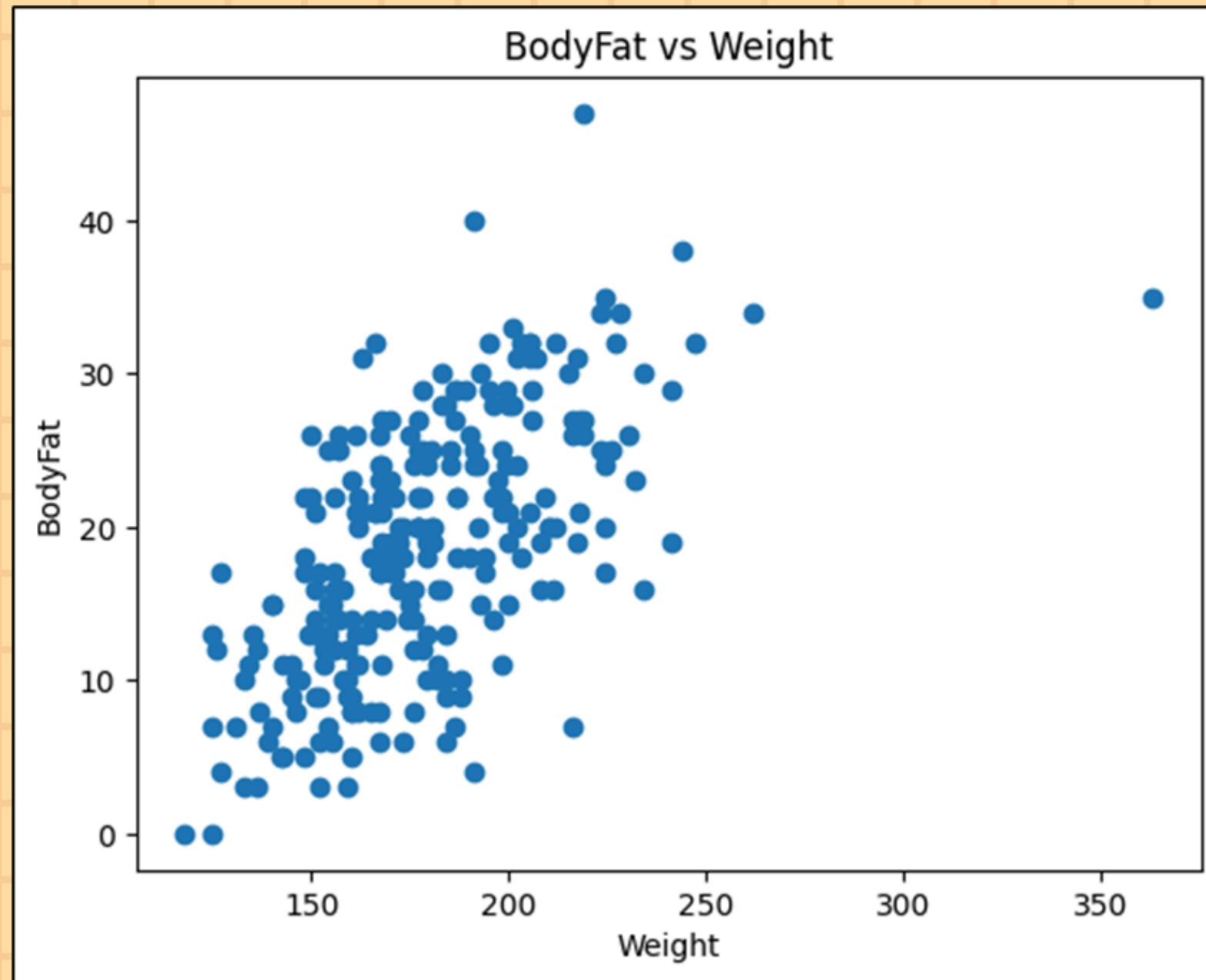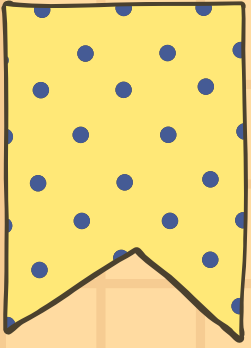
HISTOGRAM

The histogram shows a normal distribution, with a slight tilt to the left. Most of the data is between 15% and 25% with some extreme cases.
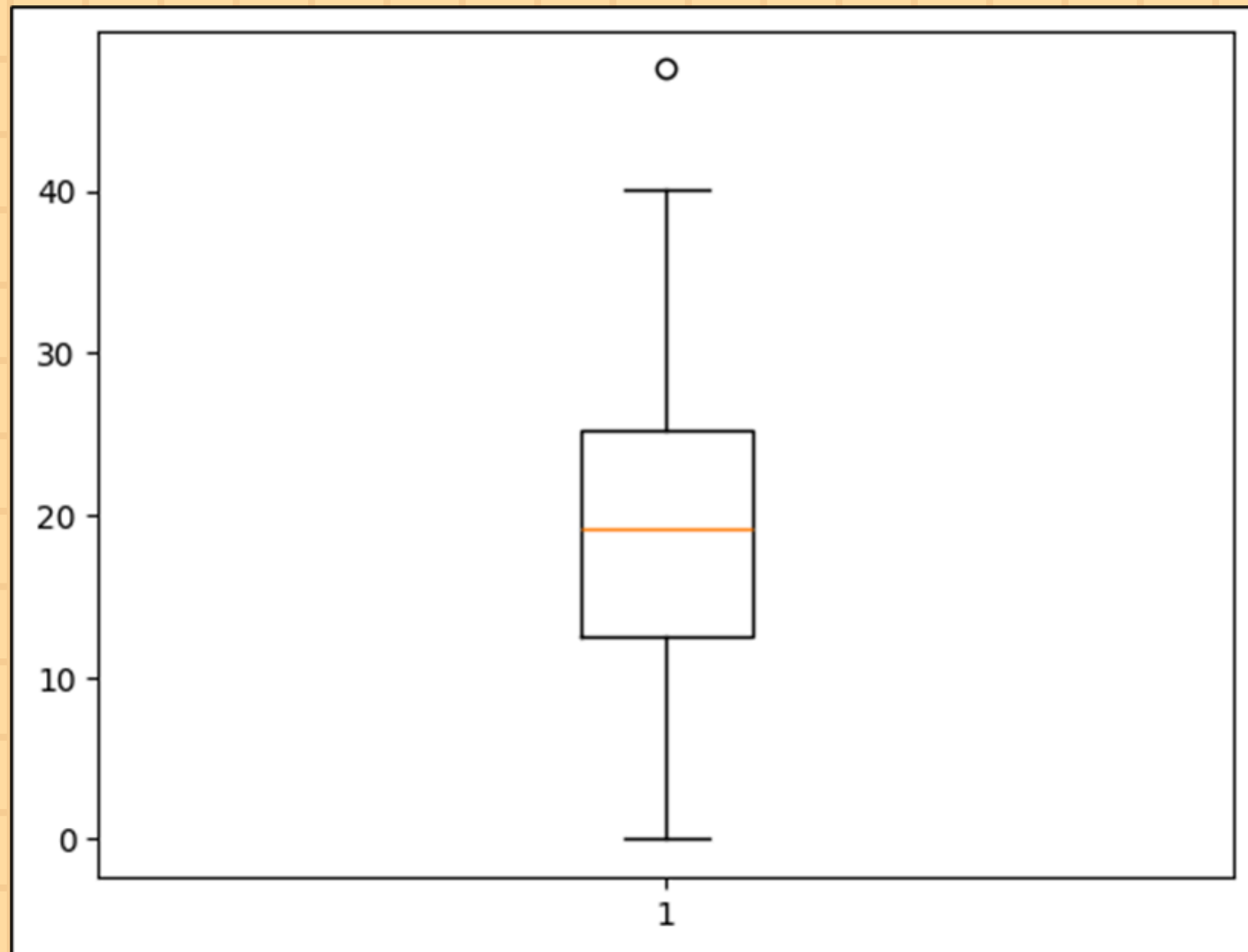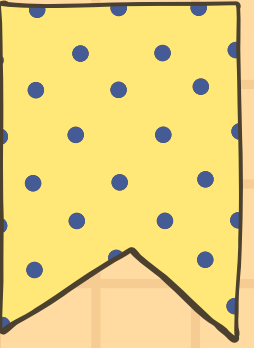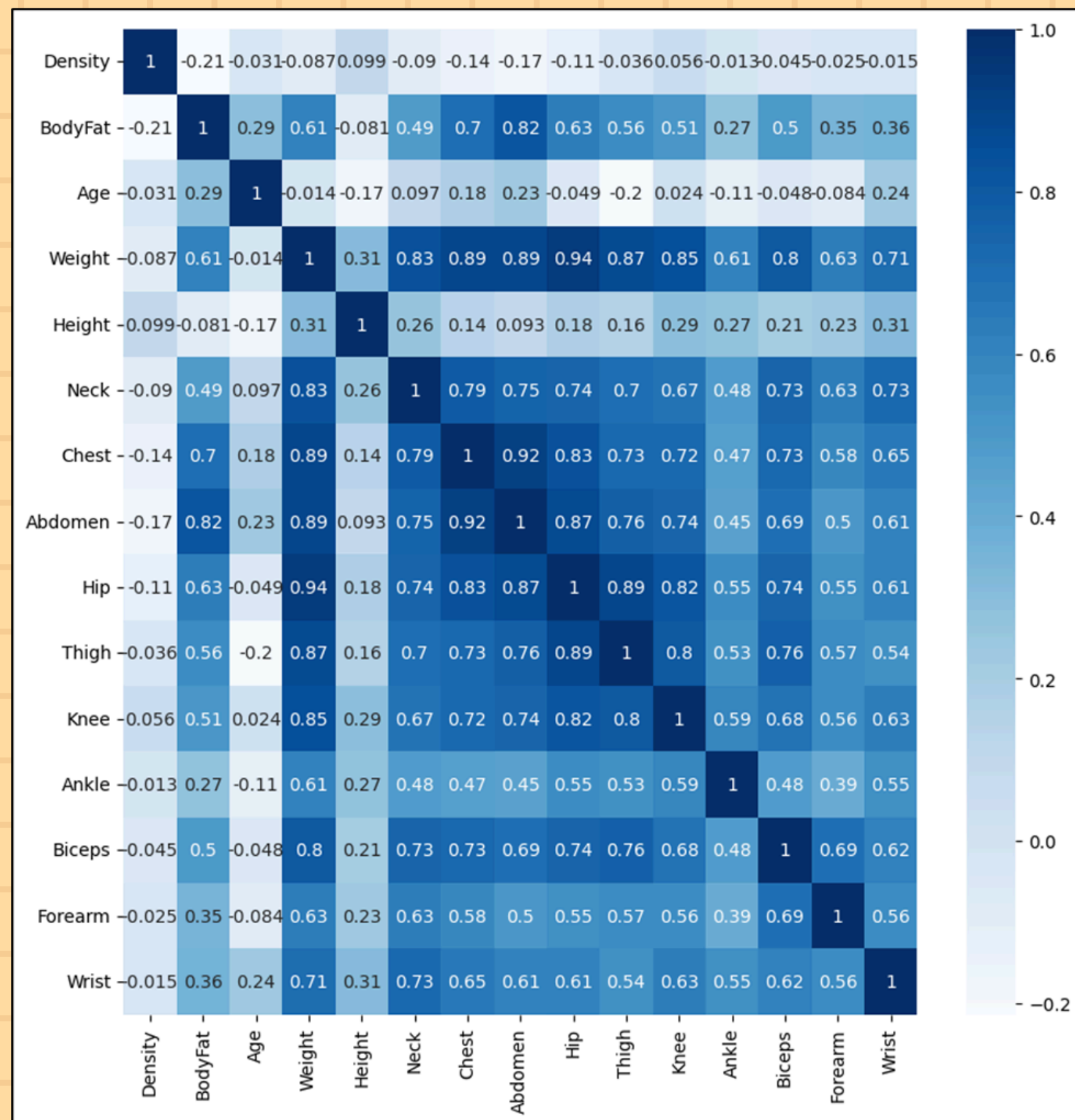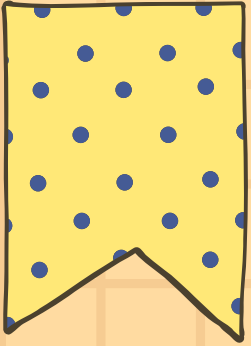
BOX PLOT

The box plot shows most body fat values are between 13 and 25 percent. The average is around 19 percent. One very high value is an outlier above 45 percent. The data is mostly balanced.
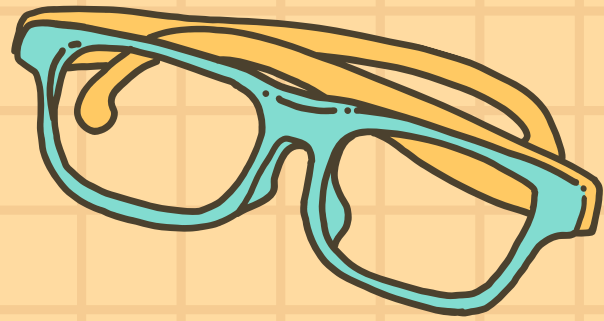
SCATTER PLOT

We can see that as the weight increases, the bodyfat increases too, and there are some outlier points. This is a positive correlation.

HEATMAP

the map shows that Abdomen and Chest have a strong relationships with BodyFat.

HYPOTHESIS TESTING

-PEARSON CORRELATION (WEIGHT VS BODYFAT): 0.610

-SPEARMAN CORRELATION (WEIGHT VS BODYFAT): 0.610

-ONE-SAMPLE T-TEST: THIS IS A ONE-SAMPLE T-TEST CHECKING IF THE AVERAGE BODY FAT IN THE DATA IS SIGNIFICANTLY DIFFERENT FROM 25. THE T-TEST STATISTIC IS -11.925 AND THE P-VALUE IS ABOUT 2.84E-26, WHICH IS MUCH LESS THAN 0.05. SINCE THE P-VALUE IS LOW, WE REJECT THE NULL HYPOTHESIS. THIS MEANS THE AVERAGE BODY FAT IN THE SAMPLE IS SIGNIFICANTLY DIFFERENT FROM 25%.
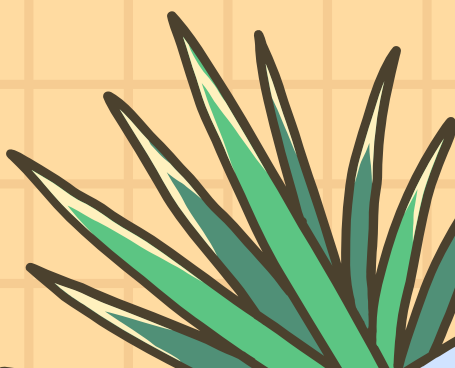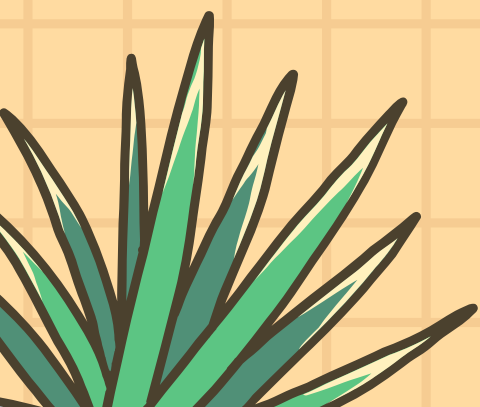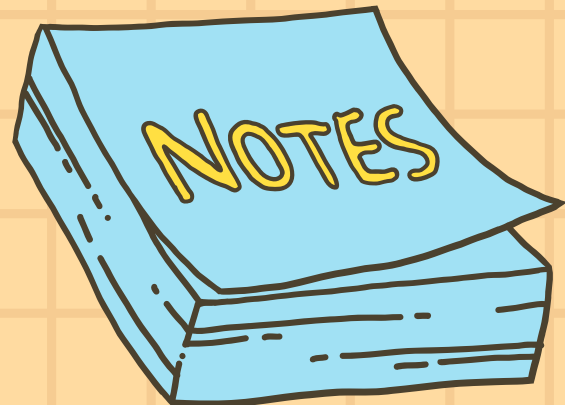
REGRESSION ANALYSIS

SIMPLE

-**Independent**: age
-**Dependent**: BodyFat
r2 Score: 0.105 (age alone isn't enough to explain bodyfat, this score basically means how good my model is)

MULTIPLE

-**Independent**: knee, ankle, biceps
-**Dependent**: BodyFat
-**r2 Score**: 0.162 (score is low likely because these 3 variables don't have a good correlation with bodyfat)

CLASSIFICATION MODELS

- LOGISTIC REGRESSION 81.5%

```
confusion matrix:
 [[54  7]
  [ 7  8]]
```

- KNN 73.6%

```
confusion matrix:
 [[47 14]
  [ 6  9]]
```
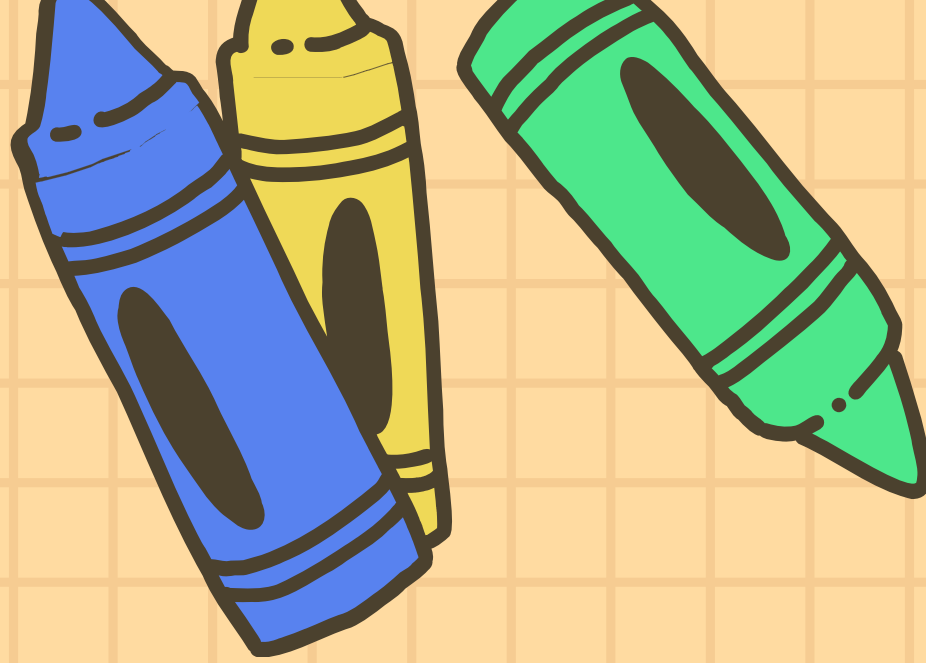
- NAIVE BAYES 80.2%

```
[[55   6]
 [ 9   6]]
```

- DECISION TREE 65.7%

```
confusion matrix:
 [[42 19]
  [ 7  8]]
```
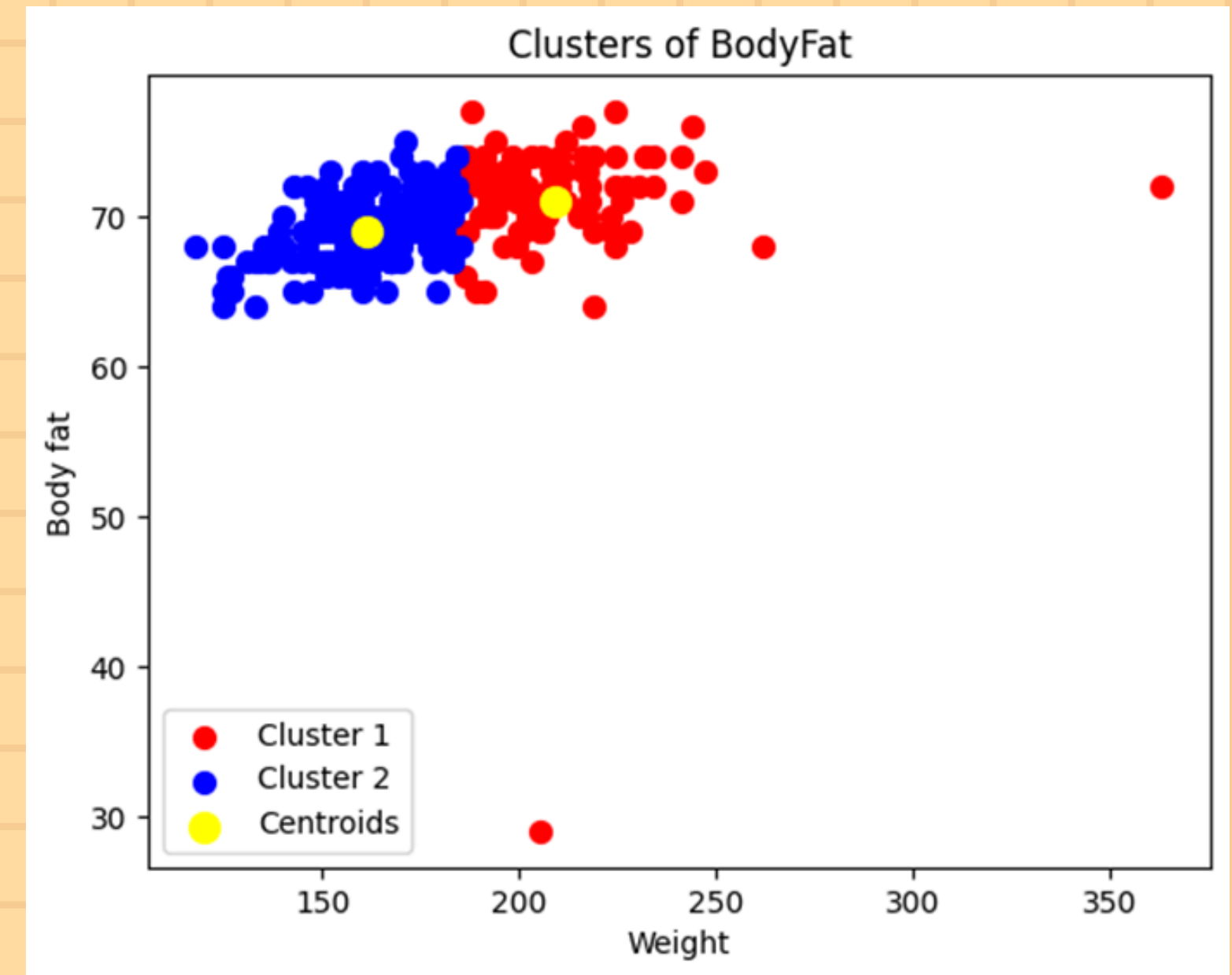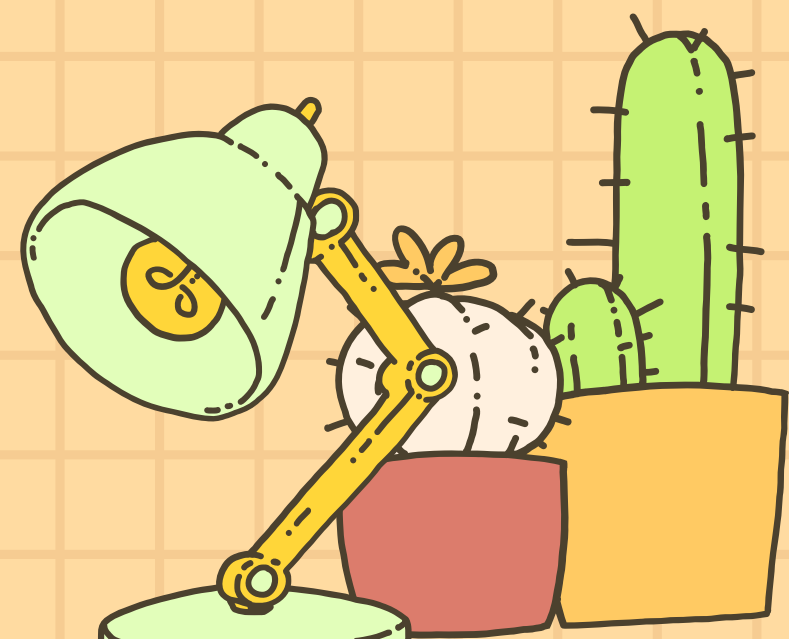
CLUSTERING

Elbow Plot

Clusters of BodyFat
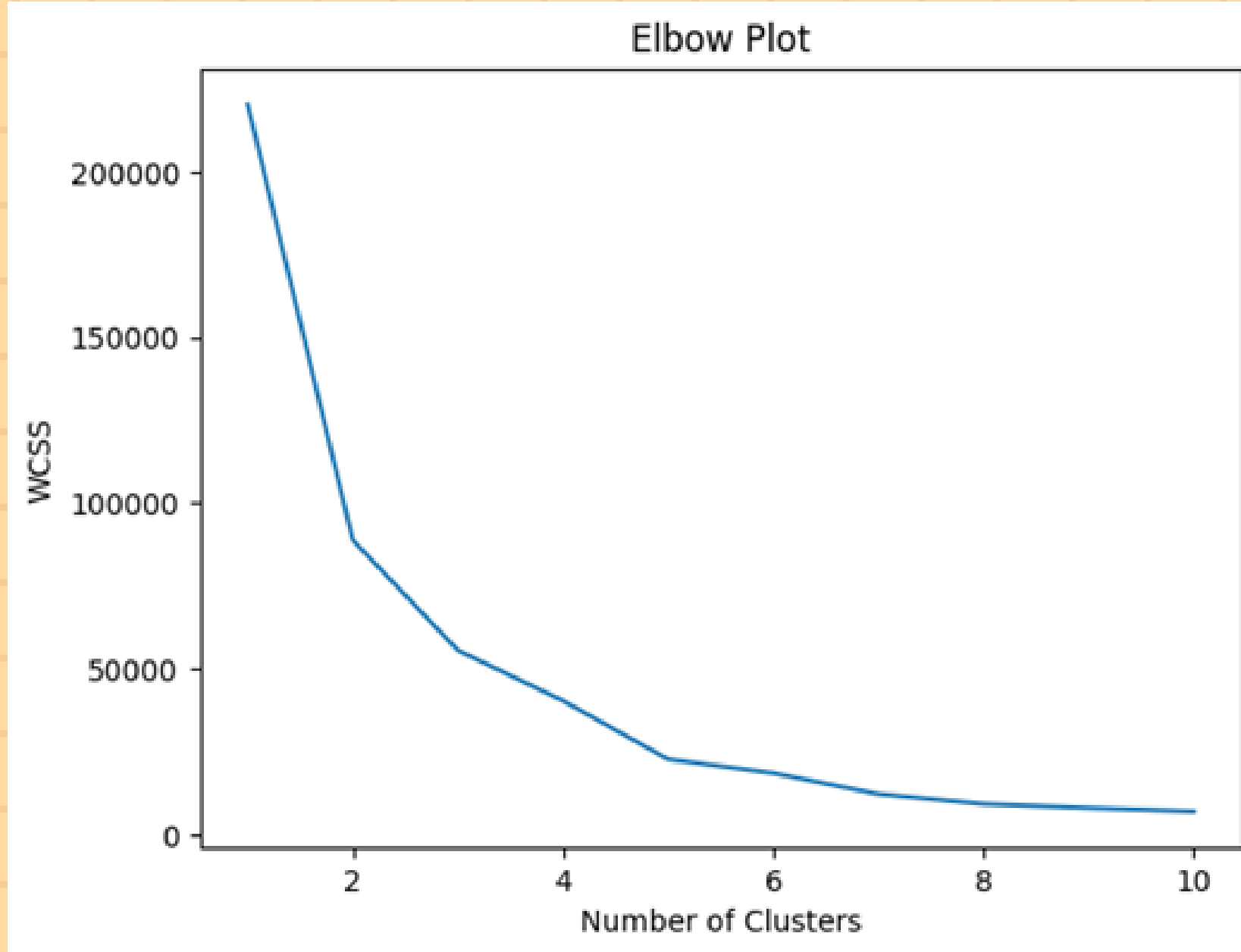
GIT VERSION CONTROL

1. Created GitHub repository.

2. Uploaded project files.

3. Invited the team members.

4. Linked the github to the google colab.

5. Evey time a change happens in the google colab it appears in GitHub and the word document files' changes can be seen when someone overrides the word file.

CONCLUSION

ABDOMEN HAS THE HIGHEST CORRELATION WITH BODYFAT

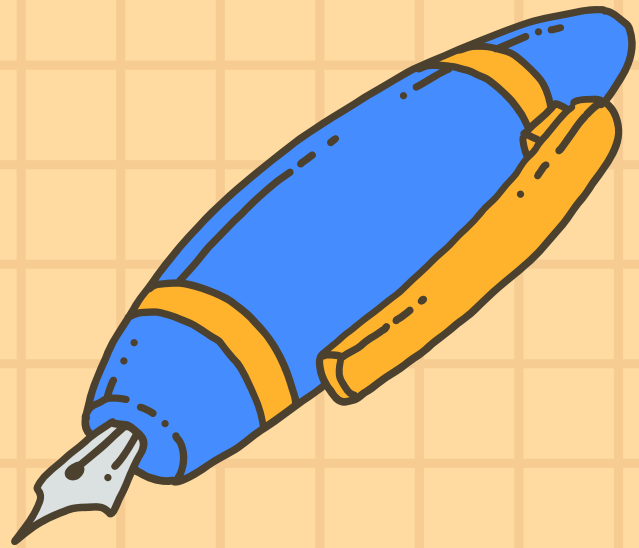MULTIPLE LINEAR REGRESSION MODEL PERFORMS BETTER THAN SIMPLE LINEAR (R2 = 0.162)

LOGISTIC REGRESSION IS THE MOST ACCURATE CLASSIFICATION MODEL (81.5%)

KMEANS CLUSTERING PROVIDED VALUABLE INSIGHT OF BODY TYPES

GITHUB WAS EFFECTIVELY USED FOR VERSION CONTROL AND TEAMWORK

COLLABORATION

WE ALL WORKED TOGETHER IN CLASS USING ASMA'S LAPTOP. WE KEPT SEARCHING FOR THE DATASET IN KAGGLE. WE TOOK TURNS WRITING CODE, FIXING ERRORS, AND TESTING. WE SAVED EVERYTHING ON GITHUB AND ASKED THE TEACHER FOR HELP WHEN NEEDED. WE ALSO WORKED ON THE REPORT TOGETHER IN CLASS. AND ASMA DID THE PRESENTATION.
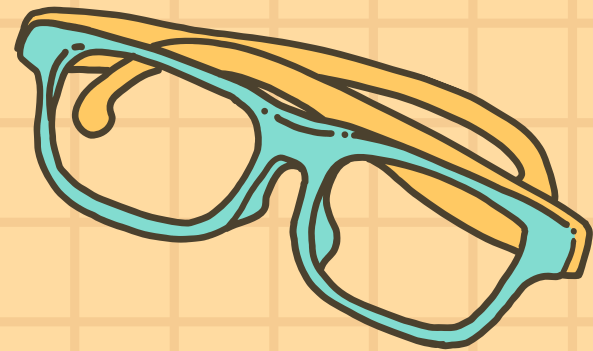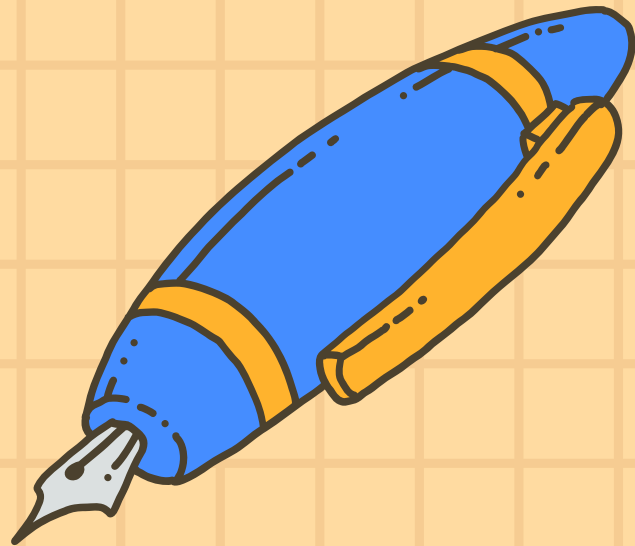
REFERENCES
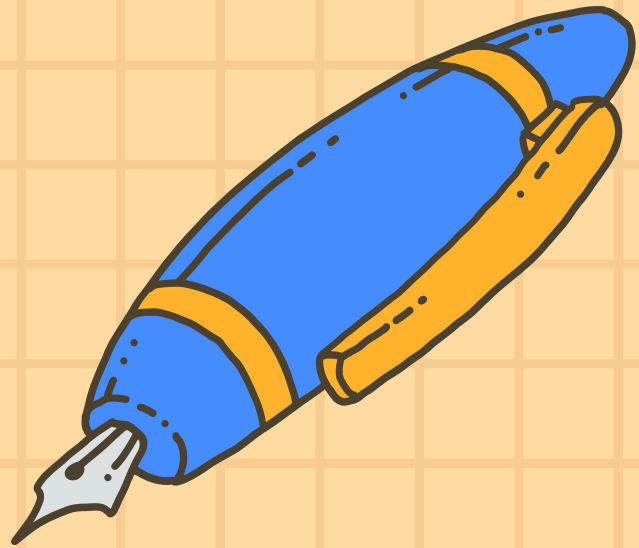
DATASET: HTTPS://WWW.KAGGLE.COM/DATASETS/FEDESORIANO/BODY-FAT-PREDICTION-DATASET
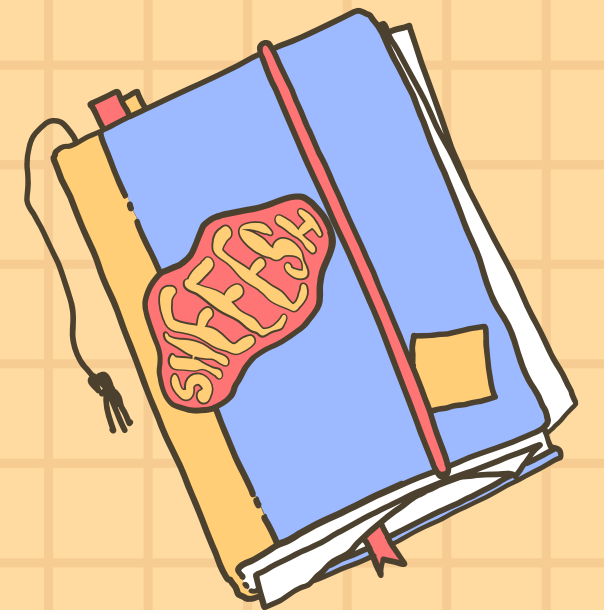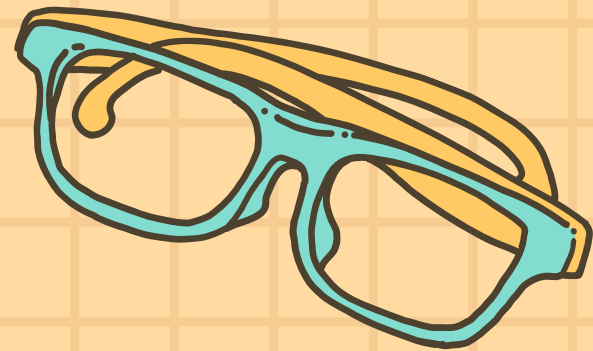
REFERENCES

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr, spearmanr, chi2_contingency, shapiro, ttest_1samp
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, recall_score, precision_score
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.cluster import KMeans
from sklearn.naive_bayes import GaussianNB
```
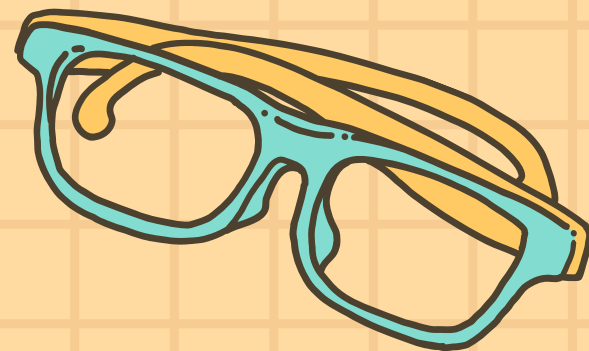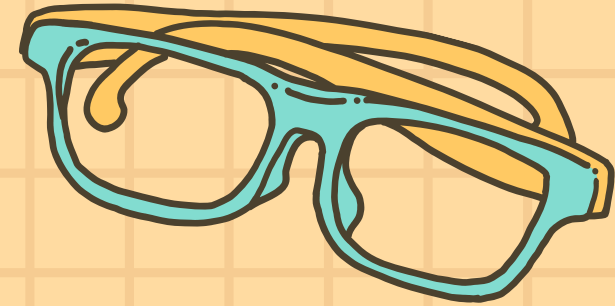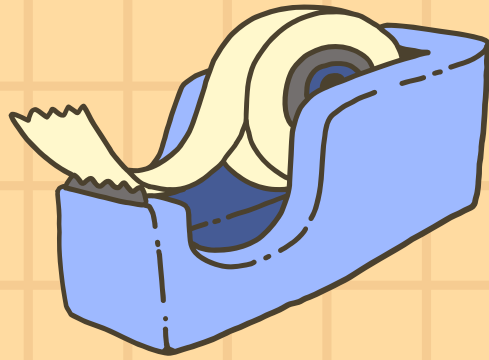
THANK YOU