**Assignment # 01**

**Data Mining**

**1.** Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).

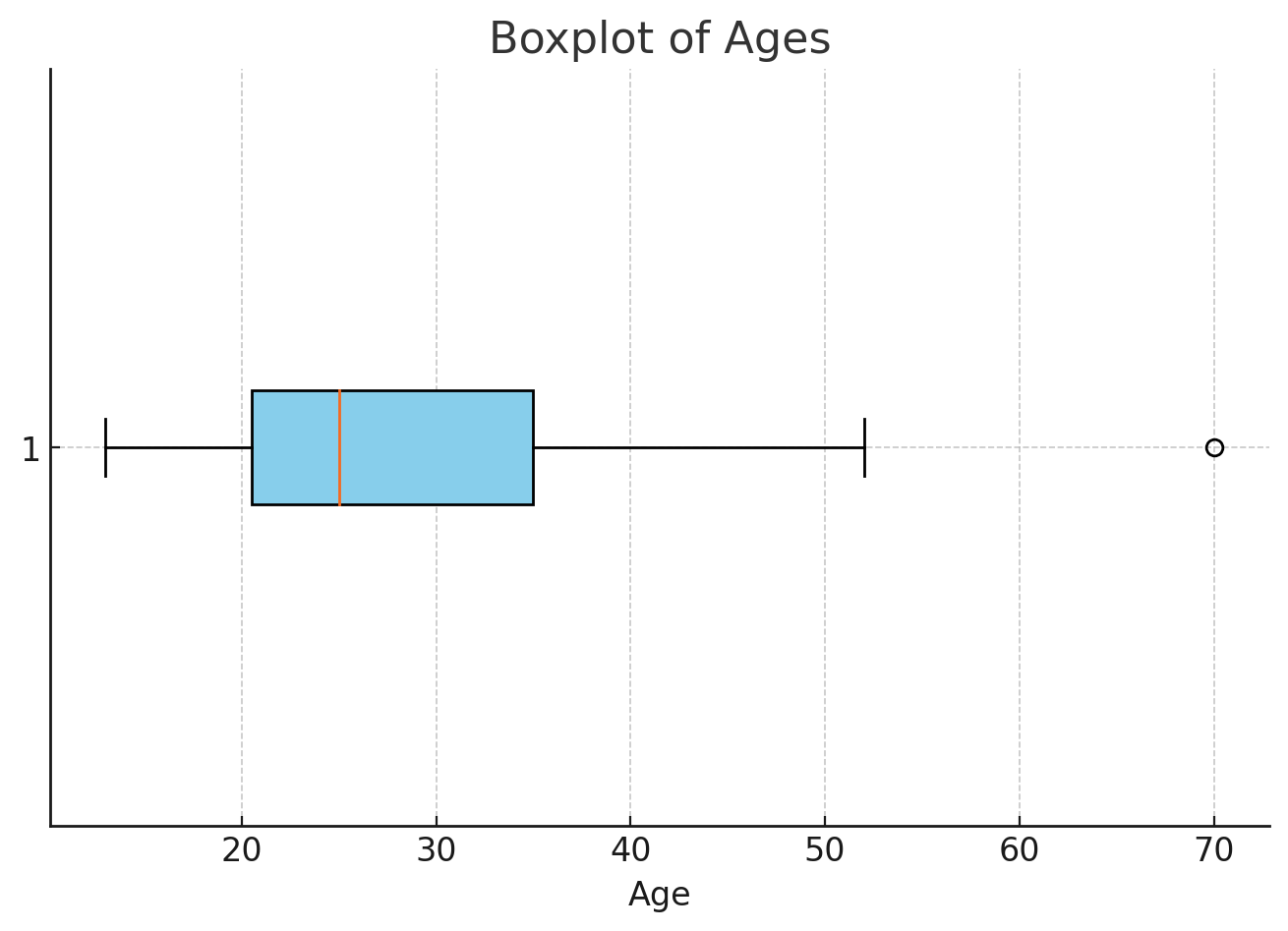
(c) What is the midrange of the data?

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

(e) Give the five-number summary of the data.

(f) Show a boxplot of the data.

(g) How is a quantile–quantile plot different from a quantile plot?



**(a) Mean and Median**

* **Mean**: 29.9629.96 (rounded to two decimal places)
* **Median**: 25.025.0

**(b) Mode and Modality**

* **Mode**: 2525 (appears most frequently)
* **Modality**: Bimodal (Two modes: 25 and 35, both appearing 4 times)

**(c) Midrange**

* **Midrange**: 41.541.5

**(d) Quartiles**

* **First Quartile (Q1)**: 20.520.5
* **Third Quartile (Q3)**: 35.035.0

**(e) Five-Number Summary**

* Minimum: 1313
* First Quartile (Q1): 20.520.5
* Median: 25.025.0
* Third Quartile (Q3): 35.035.0
* Maximum: 7070

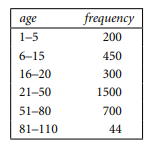
**(f) Boxplot**

The boxplot provides a visual summary of the data, including the median, quartiles, and potential outliers.

**(g) Quantile–Quantile Plot vs. Quantile Plot**

1. **Quantile Plot**:
   * Displays the data points plotted against their corresponding quantiles.
   * Helps visualize the distribution of the data.
2. **Quantile–Quantile (Q-Q) Plot**:
   * Compares the quantiles of the dataset against a theoretical distribution (e.g., normal distribution).
   * Used to check if the data follows a specific distribution.

**2.** Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:



Compute an approximate median value for the data.

To approximate the **median** for grouped data, we use the following formula:

Where:

* LL: Lower boundary of the median class
* NN: Total number of frequencies
* FF: Cumulative frequency before the median class
* fmf\_m: Frequency of the median class
* ww: Width of the median class interval

**Given Data:**

| **Age Interval** | **Frequency** |
| --- | --- |
| 1–5 | 200 |
| 6–15 | 450 |
| 16–20 | 300 |
| 21–50 | 1500 |
| 51–80 | 700 |
| 81–110 | 44 |

1. **Find NN (total frequency):**

N=200+450+300+1500+700+44=3194N = 200 + 450 + 300 + 1500 + 700 + 44 = 3194

1. **Find :**
2. **Locate the Median Class:**
   * Cumulative frequency:
     + 1–5:2001–5: 200
     + 6–15:200+450=6506–15: 200 + 450 = 650
     + 16–20:650+300=95016–20: 650 + 300 = 950
     + 21–50:950+1500=245021–50: 950 + 1500 = 2450
     + 51–80:2450+700=315051–80: 2450 + 700 = 3150
     + 81–110:3150+44=319481–110: 3150 + 44 = 3194
   * The median class is 21–5021–50, as 15971597 falls within this interval.
3. **Identify Parameters:**
   * L=21L = 21 (lower boundary of 21–5021–50)
   * F=950F = 950 (cumulative frequency before 21–5021–50)
   * fm=1500f\_m = 1500 (frequency of 21–5021–50)
   * w=50−21=29w = 50 - 21 = 29 (width of the interval)
4. **Apply the Formula:**

**Calculation:**

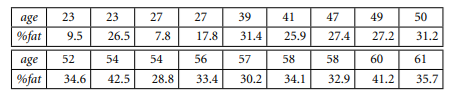
Median=21 + (0.4313 . 29)

Median= 21 + 12.51 33.51

**Approximate Median:**

The approximate median value is **33.51**.

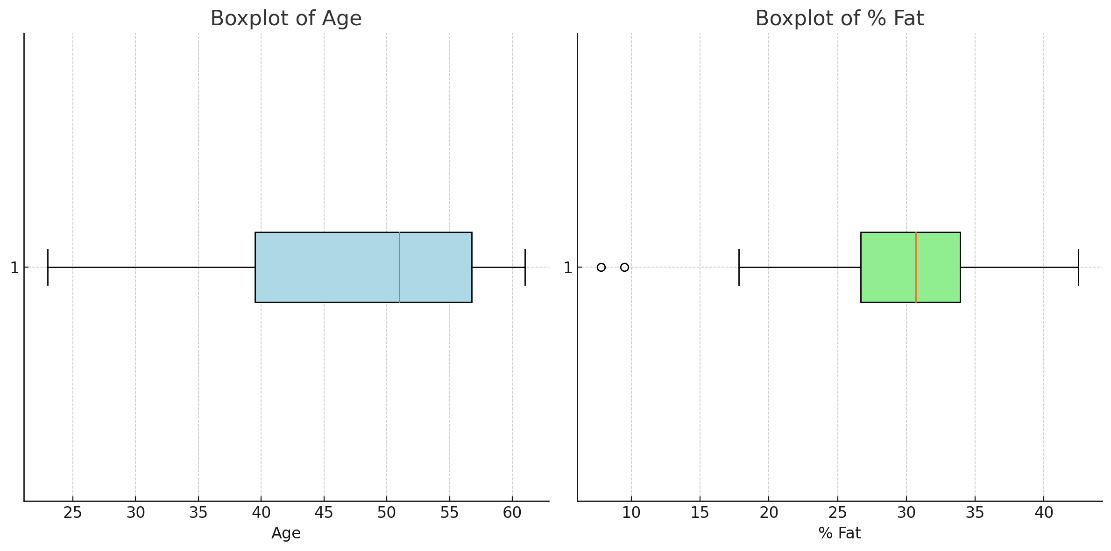
**3**. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:



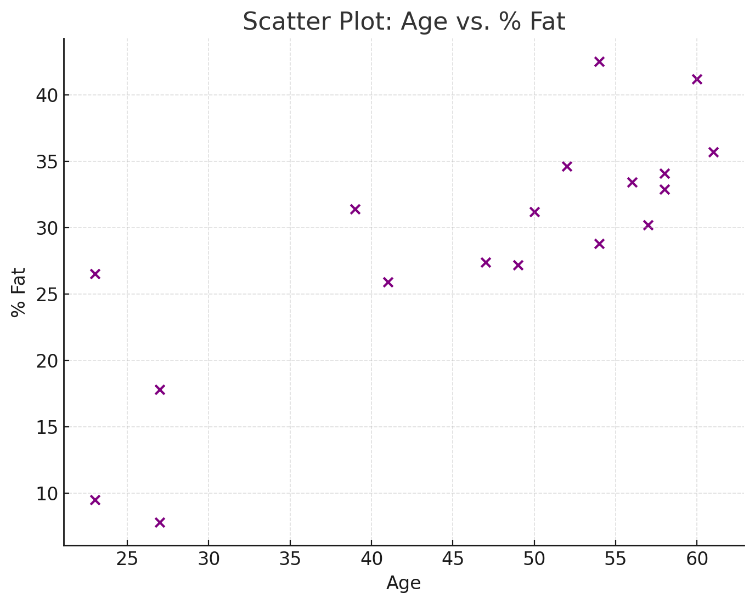
1. Calculate the mean, median, and standard deviation of age and %fat.

* **Age**:
  + Mean: 46.4446.4446.44 (rounded to two decimal places)
  + Median: 51.051.051.0
  + Standard Deviation: 13.2213.2213.22 (rounded to two decimal places)
* **% Fat**:
  + Mean: 28.7828.7828.78 (rounded to two decimal places)
  + Median: 30.730.730.7
  + Standard Deviation: 9.259.259.25 (rounded to two decimal places)

1. Draw the boxplots for age and %fat.



1. Draw a scatter plot based on these two variables.



4. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects.

(b) Compute the Manhattan distance between the two objects. (c) Compute the Minkowski distance between the two objects, using q = 3.

(d) Compute the supremum distance between the two objects.

(a) **Euclidean Distance**: 6.716.716.71 (rounded to two decimal places)  
(b) **Manhattan Distance**: 111111  
(c) **Minkowski Distance (q = 3)**: 6.156.156.15 (rounded to two decimal places)  
(d) **Supremum Distance**: 666

**5. Consider the following text (Tweets):**

@MeNyrbie @Phil\_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8

advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate supplies of regular meds but not over order

Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P

"My food stock is not the only one which is empty...

PLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need.

Stay calm, stay safe.

#COVID19france #COVID\_19 #COVID19 #coronavirus #confinement #Confinementotal #ConfinementGeneral https://t.co/zrlG0Z520j"

"Me, ready to go at supermarket during the #COVID19 outbreak.

Not because I'm paranoid, but because my food stock is litteraly empty. The #coronavirus is a serious thing, but please, don't panic. It causes shortage...

#CoronavirusFrance #restezchezvous #StayAtHome #confinement https://t.co/usmuaLq72n"

As news of the regionÂ’s first confirmed COVID-19 case came out of Sullivan County last week, people flocked to area stores to purchase cleaning supplies, hand sanitizer, food, toilet paper and other goods, @Tim\_Dodson reports https://t.co/cfXch7a2lU

Cashier at grocery store was sharing his insights on #Covid\_19 To prove his credibility he commented "I'm in Civics class so I know what I'm talking about". https://t.co/ieFDNeHgDO

**Apply Data cleaning and use TF method for Feature representation. Consider the Unigrams as features.**