Question 1

- **Gender**: can be pre-processed into three different groups: Male and female genders can be encoded as 1 and 0, respectively, while the preference to withhold gender can be encoded as 2.
- **Height**: can be pre-processed using cm after conversion. You can search for single and double quotations to get the different heights that are in Feet and Inches.
- **Earlobe**: can be pre-processed into two different groups: Attached and Detached can be encoded as 1 and 0, respectively.
- **Hair**: has the qualities Straight, Wavy, and Curly. They can each be encoded with the numbers 2, 1, or 0.
- **Music Instruments:** One-hot encoding can be used to encode additional columns with 1 for those who play that instrument and 0 for those who don't.
- **Video Games/Extracurricular/Exercise Time**: The average of the range is used for fields with ranges, and words like hours or so are eliminated.
- **Eye/Nose/Head/Hand Length**: can be converted into cm from the different units.
- **Sick Time**: transformed the string value "Once" to the number 1.

Question 2

The Manhattan distance between two points (x_1, y_1) and (x_2, y_2) in a 2D feature plane is given by:

$$|\, x_2 - x_1 \,| \; + \; |\, y_2 - y_1 \,|$$

The Euclidean distance between two points $(x1, y1)$ and $(x2, y2)$ in a 2D feature plane is given by:

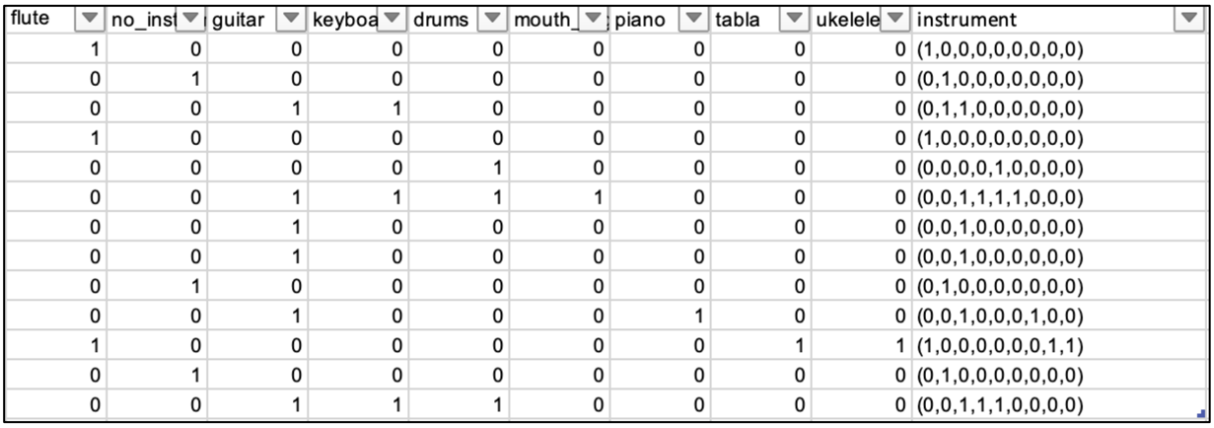
$$\sqrt[2]{(x_2 - x_1)^2 + (y_2 - y_1\,)^2}$$

For the Manhattan distance and Euclidean distance to be the same for two points in a 2D feature plane, the following condition must hold:

$$|\, x_2 - x_1 \,| \; + \; |\, y_2 - y_1 \,| = \sqrt[2]{(x_2 - x_1)^2 + (y_2 - y_1\,)^2}$$

The two points must be situated on a straight line with a slope of either 1 or -1, passing through the origin, for this condition to be true. In other words, the two points have an x-coordinate difference that is equal to their y-coordinate difference or vice versa.

ONE-HOT ENCODING

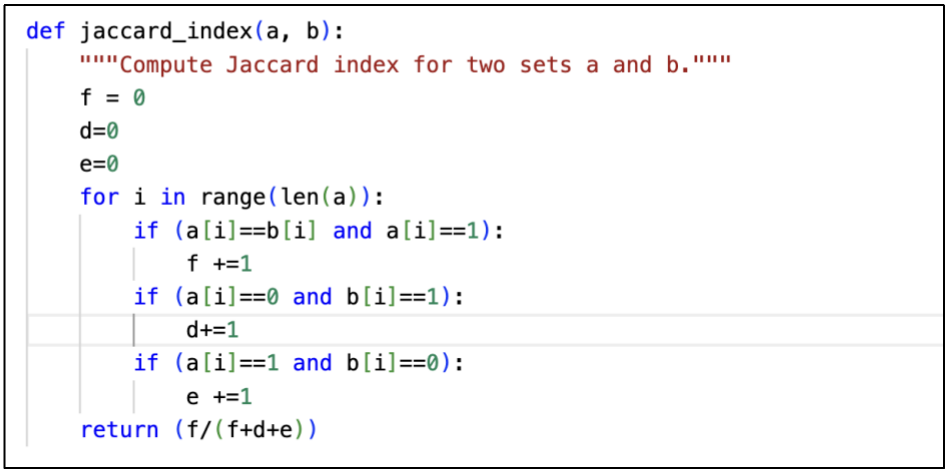


flute	no_inst	guitar	keyboa	drums	mouth_	piano	tabla	ukelele	instrument
1	0	0	0	0	0	0	0	0	(1,0,0,0,0,0,0,0,0)
0	1	0	0	0	0	0	0	0	(0,1,0,0,0,0,0,0,0)
0	0	1	1	0	0	0	0	0	(0,1,1,0,0,0,0,0,0)
1	0	0	0	0	0	0	0	0	(1,0,0,0,0,0,0,0,0)
0	0	0	0	1	0	0	0	0	(0,0,0,0,1,0,0,0,0)
0	0	1	1	1	1	0	0	0	(0,0,1,1,1,1,0,0,0)
0	0	1	0	0	0	0	0	0	(0,0,1,0,0,0,0,0,0)
0	0	1	0	0	0	0	0	0	(0,0,1,0,0,0,0,0,0)
0	1	0	0	0	0	0	0	0	(0,1,0,0,0,0,0,0,0)
0	0	1	0	0	0	1	0	0	(0,0,1,0,0,0,1,0,0)
1	0	0	0	0	0	0	1	1	(1,0,0,0,0,0,0,1,1)
0	1	0	0	0	0	0	0	0	(0,1,0,0,0,0,0,0,0)
0	0	1	1	1	0	0	0	0	(0,0,1,1,1,0,0,0,0)

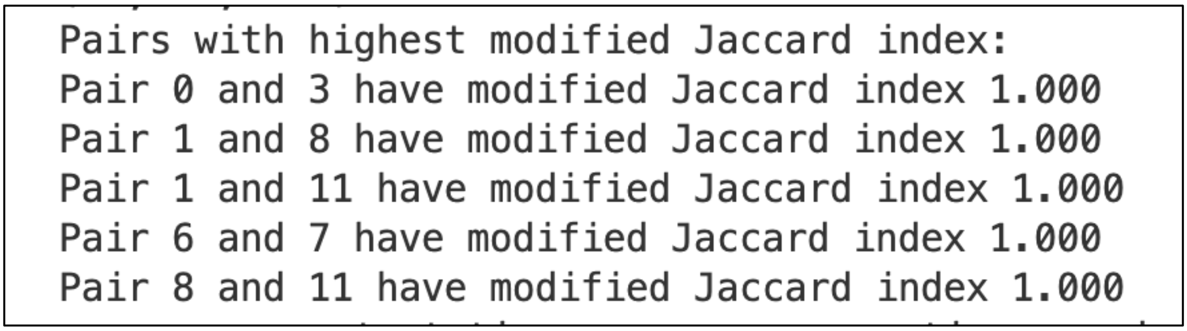
Note: In the above table, 1 denotes that the instance can play that particular instrument and 0 denotes that they cannot.

QUESTION 3

- The Jaccard index is a measure of similarity between two sets, defined as the size of the intersection divided by the size of the union of the sets. In the context of this problem, we can use the Jaccard index to measure the similarity between two instances based on the musical instruments they can play. To calculate the Jaccard index for two instances, we first represent each instance as a set of musical instruments that the person can play. We can then calculate the size of the intersection and union of the two sets, and divide the size of the intersection by the size of the union to get the Jaccard index.

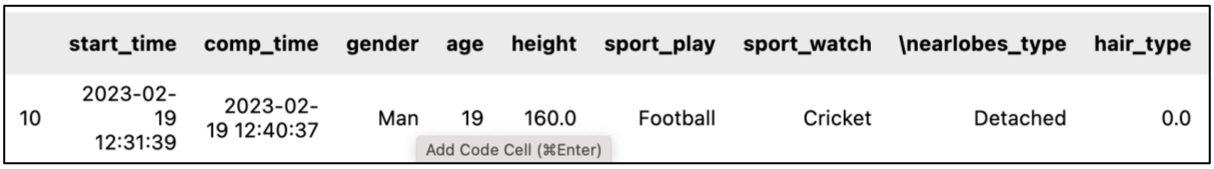


```python
def jaccard_index(a, b):
    """Compute Jaccard index for two sets a and b."""
    f = 0
    d=0
    e=0
    for i in range(len(a)):
        if (a[i]==b[i] and a[i]==1):
            f +=1
        if (a[i]==0 and b[i]==1):
            d+=1
        if (a[i]==1 and b[i]==0):
            e +=1
    return (f/(f+d+e))
```



```
Pairs with highest modified Jaccard index:
Pair 0 and 3 have modified Jaccard index 1.000
Pair 1 and 8 have modified Jaccard index 1.000
Pair 1 and 11 have modified Jaccard index 1.000
Pair 6 and 7 have modified Jaccard index 1.000
Pair 8 and 11 have modified Jaccard index 1.000
```

QUESTION 4

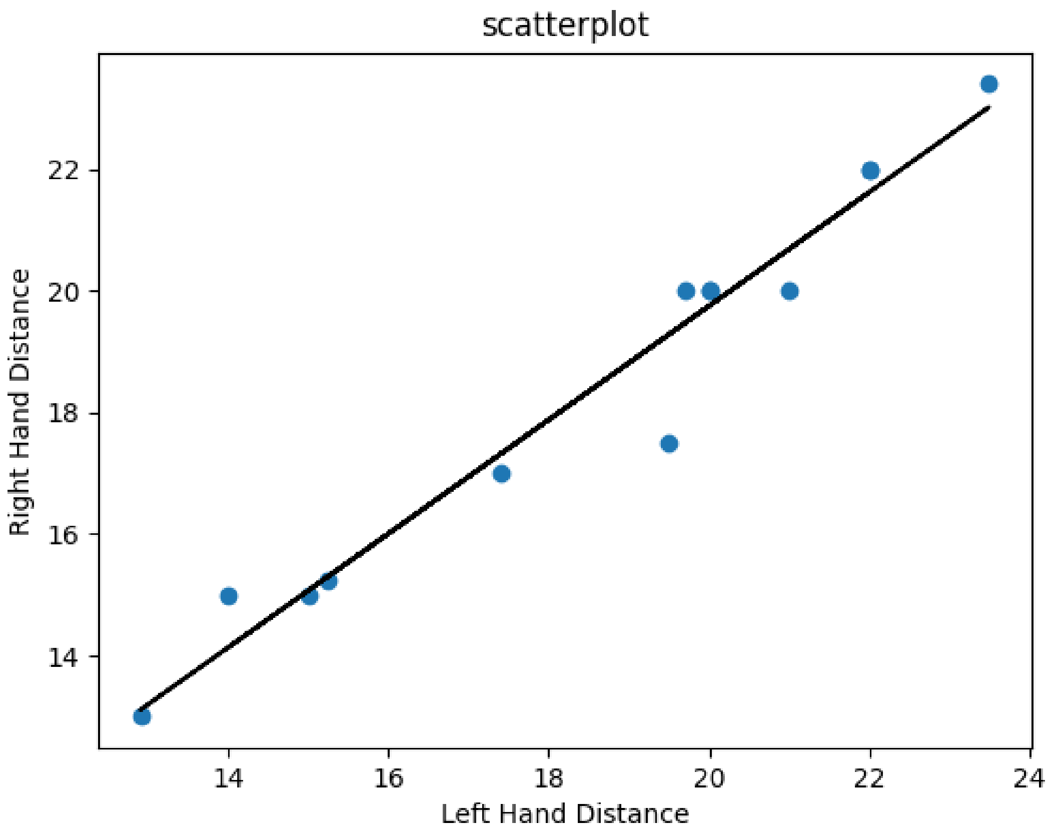


	start_time	comp_time	gender	age	height	sport_play	sport_watch	\nearlobes_type	hair_type
10	2023-02-19 12:31:39	2023-02-19 12:40:37	Man	19	160.0	Football	Cricket	Detached	0.0

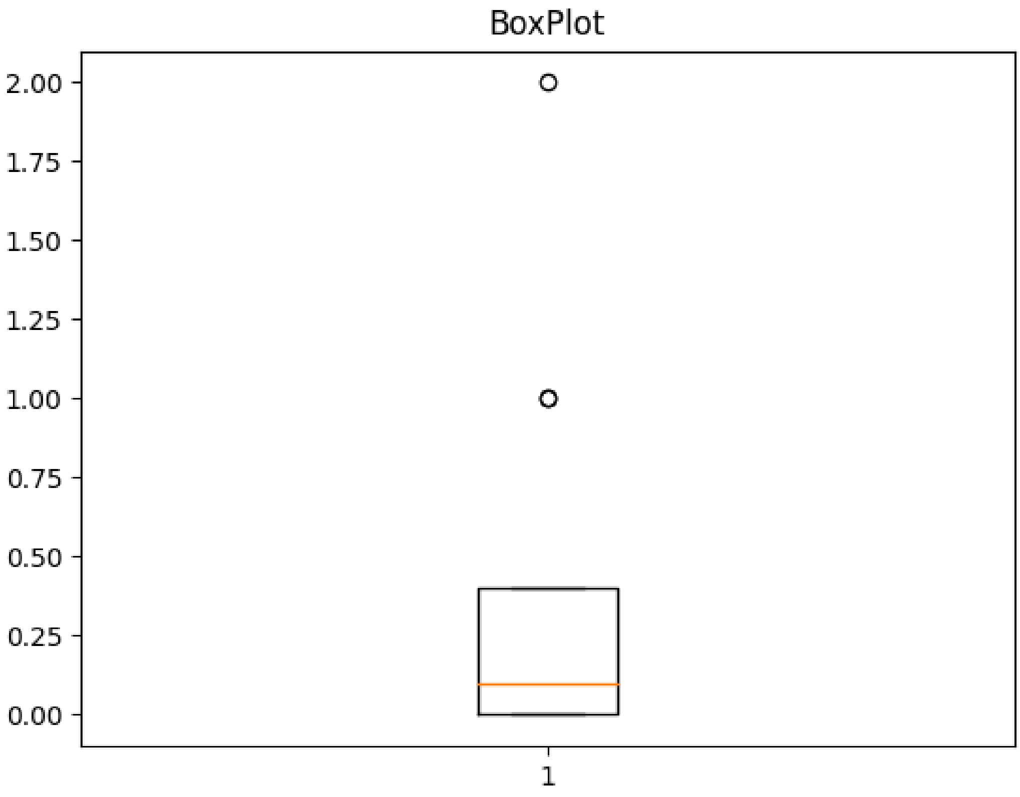
Add Code Cell (⌘Enter)

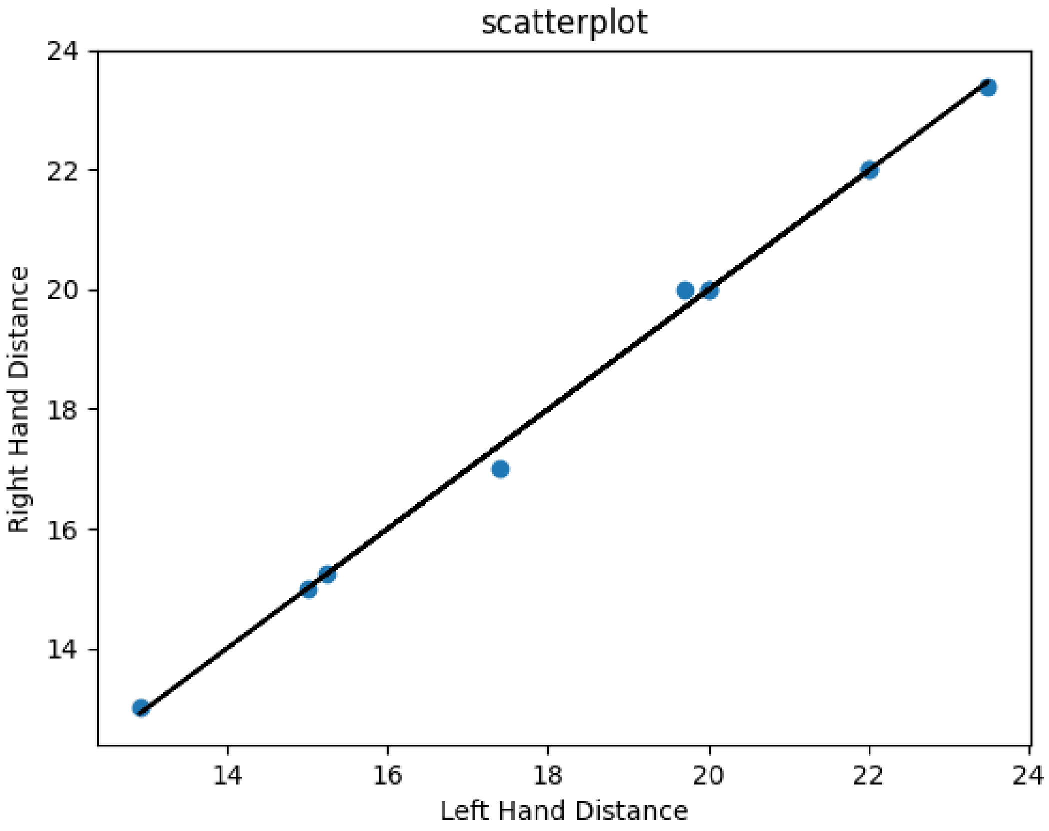
Hair Type – Curly
Serial Number -11(as it starts from 0)

Question 5



6.499 %-slope change





Slope Change- 0.211 %

As we drop the outliers, we see that slope of the regression lines comes very close to 1, our expected slope, which confirms our assumption that the length of hand should be equal.