Name: Malik Majette(mamajett), Qua Jones(qyjones), Wenting Zheng(wzheng8),

1.

1. Hair color (Black, Blonde, Red): Nominal | Discrete
2. Level of agreement (yes, maybe, no): Ordinal | Discrete
3. Income earned in a week: Ratio | Discrete
4. Celsius temperature: Interval | Continuous
5. Genotype (Bb, bb, BB, bB): Nominal, under the assumption there is no ranking to a person’s genotype | Discrete
6. ISBN numbers for books: Nominal | Discrete
7. Time in terms of AM or PM: Interval | Binary
8. Waiting number for restaurant: Ordinal | Discrete
9. Years of work experience: Ratio | Discrete
10. Categorization of clothing (hat, shirt, pants, shoes): Nominal | Discrete
11. Angles as measured in degrees between 0 and 360: Ratio | Continuous
12. Ratings of movies (G, PG, R): Ordinal | Discrete
13. Coat check number. Nominal, under the assumption that coat check numbers are not determined by type of coat and do not influence order of retrieving coat | Discrete

2.

1. What are the maximum and minimum values of tf’ij and tf’’ij respectively? Please specify what cases the max and min value achieves.
2. tf’ij maximum value: The maximum value of tf’ij is 1 and is approached as the term frequency in document *j* increases and the term frequency in the while collection decreases.
3. tf’ij minimum value: The minimum value of tf’ij is 0 and is approached as a term becomes more common in the collection of documents and less frequent in document *j*.
4. tf’’ij maximum value is positive infinity when the sum of dk from 1 to dfi is its minimum, the sum of dk from 1 to m is its maximum, and tfij is its maximum.
5. tf’’ij minimum value is 0 and is approached as tfij = 1, sum of dk from 1 to m equals to the sum of dk from 1 to dk. In the other words, all the documents has ith word and the word only occurs once in jth document

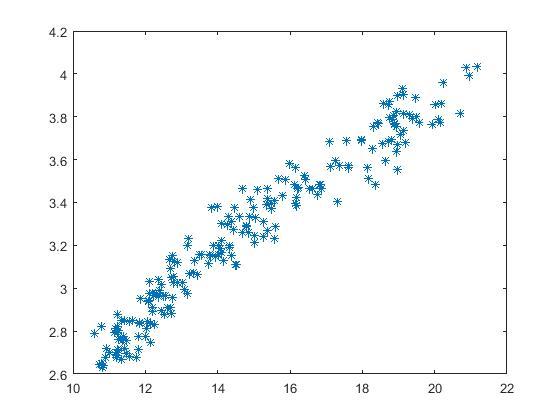
3.

1. Stratified sampling because we want to make sure we select enough data from the patients who have albinism. We will randomly select 50 albinotic patients and 950 normal skin patients.
2. If there is some relationship between hypertension and albinism, the recorded data can be a noise.

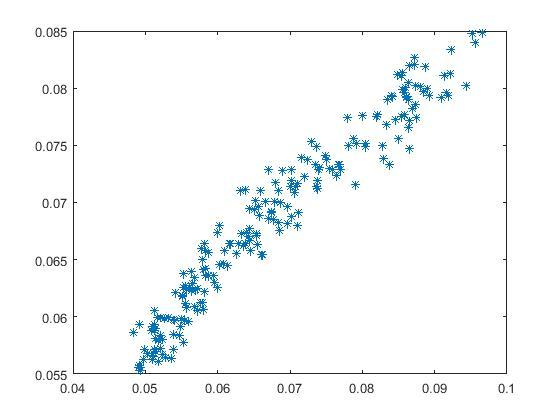
4. Result:

(a)  
[[ 1. 0. 0. 0. 0.]  
 [ 0. 1. 0. 0. 0.]  
 [ 0. 0. 1. 0. 0.]  
 [ 0. 0. 0. 1. 0.]  
 [ 0. 0. 0. 0. 1.]]  
(b)  
[[ 1. 3. 0. 0. 0.]  
 [ 0. 3. 0. 0. 0.]  
 [ 0. 3. 1. 0. 0.]  
 [ 0. 3. 0. 1. 0.]  
 [ 0. 3. 0. 0. 1.]]  
(c)   
19.0  
(d)  
[[ 1. 0. 0. 0. 0.]  
 [ 3. 3. 3. 3. 3.]  
 [ 0. 0. 1. 0. 0.]  
 [ 0. 0. 0. 1. 0.]  
 [ 0. 0. 0. 0. 1.]]  
(e)  
7.0  
(f)  
[[ 7.00665128 4.11243332 3.67224527 2.11455177 4.94422009]  
 [ 4.46001369 4.66022491 8.01774712 3.73689752 7.18145492]  
 [ 5.39237172 6.41621646 10.60856693 4.91206461 5.03864033]  
 [ 8.18106358 4.63705237 3.63705706 4.65840079 6.27522943]  
 [ 5.38609745 4.83089781 6.43875362 4.78055199 4.68329528]]  
(g)  
[[ 31.24976066 19.16486421 29.4431339 7.90186329 35.50669372]  
 [ 8.18733785 6.22237102 7.80687037 4.78991341 6.63057448]]  
(h)  
[[ 62.49952132 57.49459263 117.77253561 39.50931644 213.04016231]  
 [ 16.3746757 18.66711306 31.2274815 23.94956704 39.78344687]]  
(i) covariance matrix =   
[[ 6.66666667 -3.33333333 6.66666667]  
 [-3.33333333 1.66666667 -3.33333333]  
 [ 6.66666667 -3.33333333 6.66666667]]  
(j)  
True

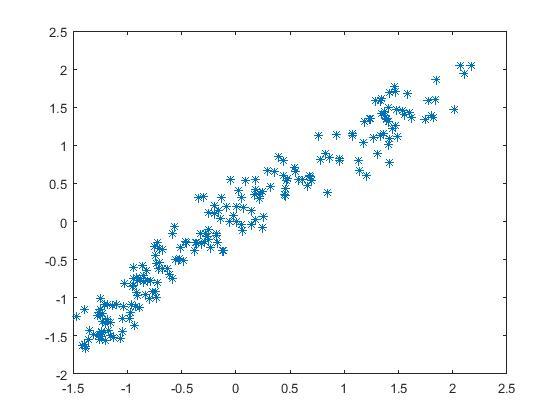
5.

1. Range = 0.0483 0.0295
   1. 

plot 1: area\_A as x, kernel\_width as y



plot 2: normalized data

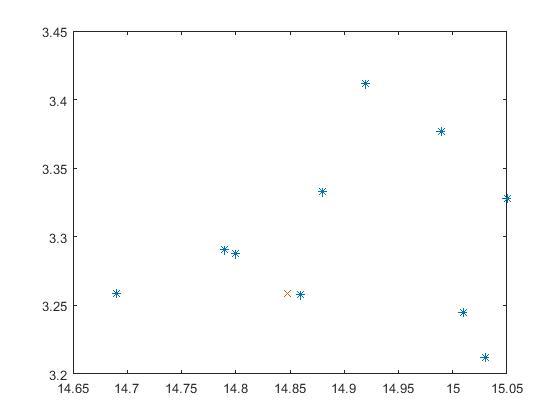


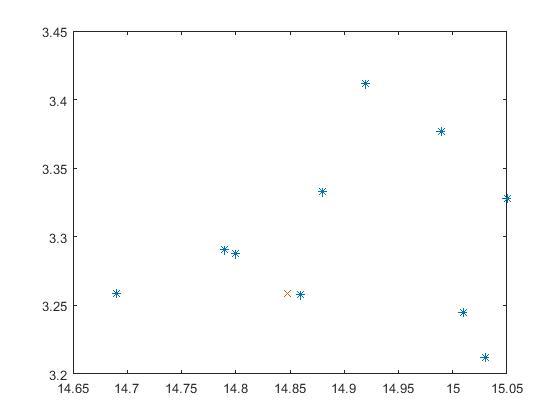
plot 3: standardized data

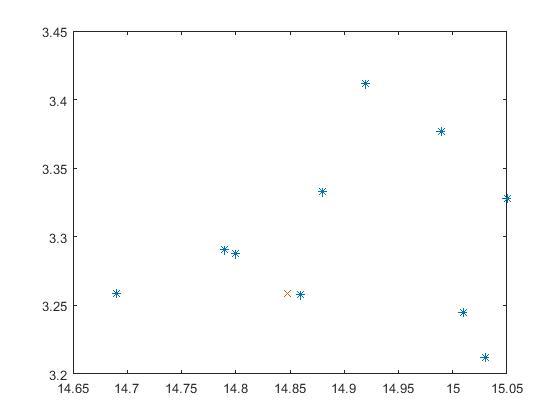
These three plots shows the positive correlation between area\_A and kernel\_width

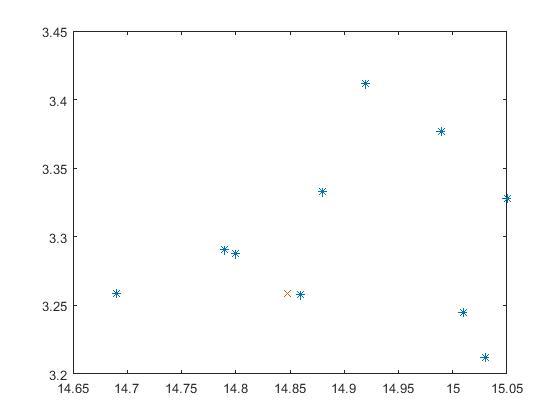
* 1. P = [14.848, 5.629]
  2. See the result computed in the code

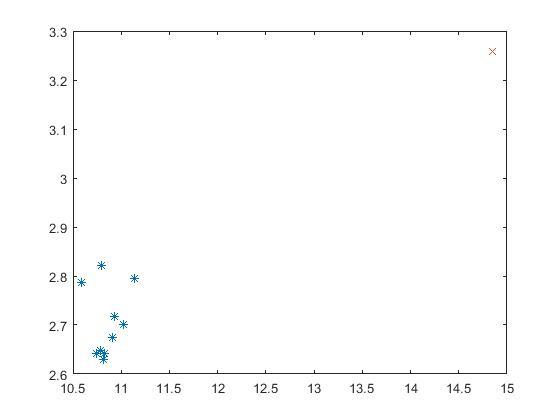
Iv & v.

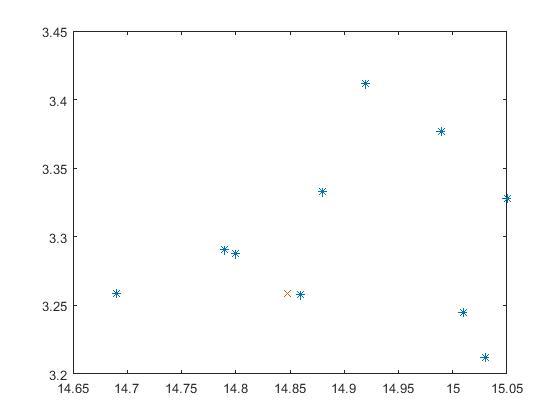
(1)Euclidean

(3)City block

(4)Minkowski

(5)Chebyshev

(6)Cosine

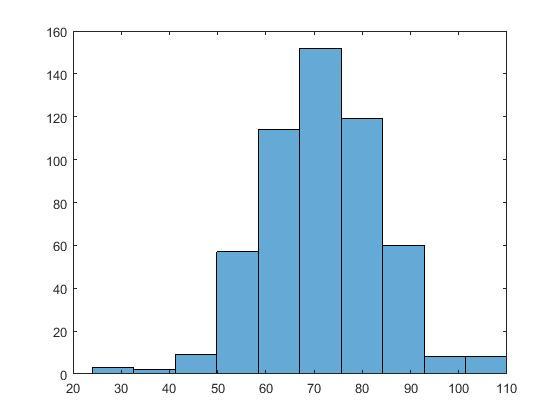
(7)Canberra

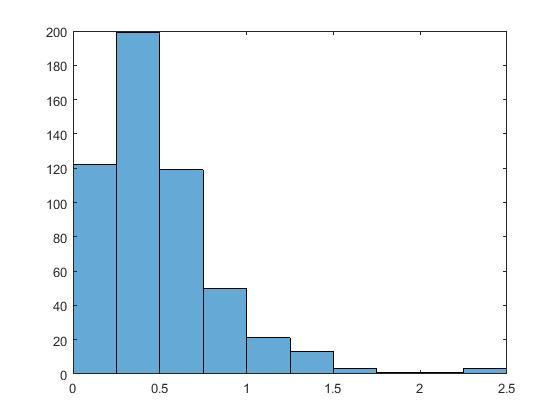
vi. Euclidean, City block, Minkowski, Chebyshev, and Canberra plots are the same. The cosine measures the cosine of the angular between the 10 closest points, so most of the points are in a line except for some outliers.

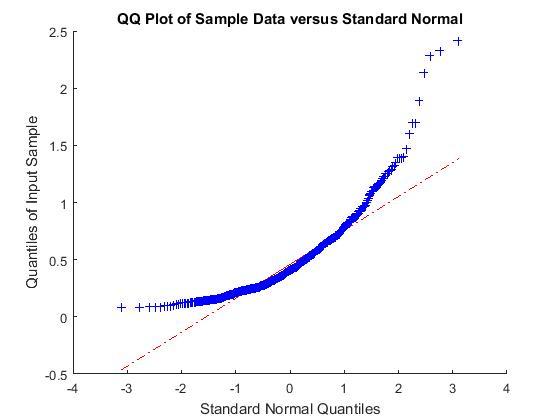
vii. Data transformation is a good way to transfer non-linear or asymmetric data to be more linear or symmetric data so that it’s easier to model. It will have less outliers and less extreme outliers.

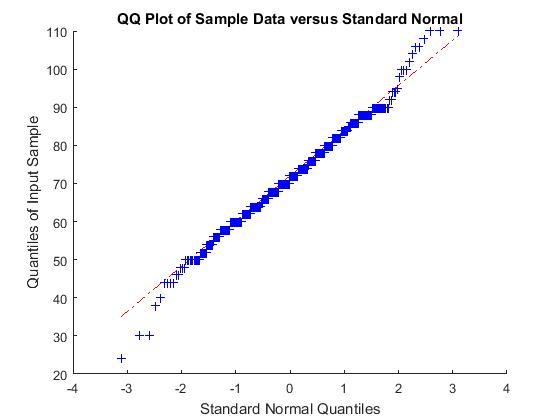
6.

1. 268 patients are diabetic and 500 are nondiabetic
2. Glucose missing rate = 0.0065, blood pressure missing rate = 0.0456, skin thickness missing rate = 0.2956, BMI missing rate = 0.0143, diabete predigree function missing rate = 0, age missing rate = 0
3. 4 ways to handle missing values: eliminate data objects; eliminate missing values; ignore the missing values during analysis; replace with all possible values(by their weight)
4. 177 patients are diabetic and 355 are nondiabetic
5. Mean = 121.0301 71.5056 29.1823 32.8902 0.5030 31.6147  
   Median = 115.0000 72.0000 29.0000 32.8000 0.4160 28.0000  
   Standard deviation = 30.9992 12.3103 10.5239 6.8811 0.3445 10.7616  
   Range = 143.0000 86.0000 92.0000 48.9000 2.3350 60.0000  
   25th = 98.5000 64.0000 22.0000 27.8500 0.2585 23.0000  
   50th = 115.0000 72.0000 29.0000 32.8000 0.4160 28.0000  
   75th = 141.5000 80.0000 36.0000 36.9000 0.6590 38.0000
6. BP = BloodPressure, DPF = DiabetesPedigreeFunction

BP histogram plot

DPF histogram plot

1. DPF quantile-quantile plot

BP quantile-quantile plot