**Q1.**

(13 points) [Yuan Zhang] Classify the following attributes as binary, discrete, or con-

tinuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval

or ratio). Some case may have more than one interpretation, so brie

y indicate your

reasoning if you think there may be some ambiguity. (each question 1 point).

(a) (1 point) Time in terms of AM or PM.

(b) (1 point) Brightness as measured by a light meter.

(c) (1 point) Brightness as measured by people's judgements.

(d) (1 point) Temperature as measured in Fahrenheit.

(e) (1 point) Bronze, Silver, and Gold medals awarded at the Olympics

(f) (1 point) Height above sea level.

(g) (1 point) Number of patients in a hospital.

(h) (1 point) ISBN numbers for books. (Look up the format on the Web.)

(i) (1 point) Ability to pass light in the terms of the following values: opaque, translu-

cent, transparent.

(j) (1 point) Military rank

(k) (1 point) Latitude/Longitude values.

(l) (1 point) Volume of a substance in cubic centimeter.

(m) (1 point) Coat check number. (When you attend an event, you can often give your

coat to someone who, in turn, gives you a numb that you can use to claim your

coat when you leave.)

**Sol.**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Binary /Discrete/ Continuous** | **Reasoning** |
| Time in terms of AM and PM | Discrete | Time has finite values from 0 to 12 |
| Brightness as measured by a light meter | Continuous | Measuring of light can be decimal values and have long range |
| Brightness as measured by people's judgments | Binary | By taking dark as `1` and bright as `0` |
| Temperature as measured in Fahrenheit | Continuous | Can have decimal values and can rise till large values (in Billions) |
| Bronze, Silver, and Gold medals awarded at the Olympics | Discrete | Can have 3 values |
| Height above sea level | Continuous | Can have decimal values and can rise till large values |
| Number of patients in a hospital | Discrete | Have numerical values |
| ISBN numbers for books | Discrete | Fixed range with numerical values |
| Ability to pass light in the terms of the following values | Discrete | Have only 3 values (Opaque-0, Translucent-1, Transparent-2) |
| Military rank | Discrete | Have numerical values |
| Latitude/Longitude values | Continuous | Have a fixed range of values |
| Volume of a substance in cubic centimeter | Continuous | Can have decimal values and no limit |
| Coat check number | Discrete | Have numerical values |

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Qualitative /Quantitative** | **Reasoning** |
| Time in terms of AM and PM | Quantitative (Interval) | There is no `0` time and time can be greater or less than |
| Brightness as measured by a light meter | Quantitative (Ratio) | Measurements can have 0 values and can be nominal, ordinal, interval and ratio |
| Brightness as measured by people's judgments | Qualitative (Nominal) | Have only two values 0 or 1 as we have taken dark and bright |
| Temperature as measured in Fahrenheit | Quantitative (Interval) | There is no `0` temperature in Fahrenheit and we can’t judge half or double hot/cold |
| Bronze, Silver, and Gold medals awarded at the Olympics | Quantitative (Ratio) | If we take number of medals awarded for each category than it can be nominal, ordinal, interval and ratio |
| Height above sea level | Quantitative (Ratio) | Measurements can have 0 values and can be nominal, ordinal, interval and ratio |
| Number of patients in a hospital | Quantitative (Ratio) | Number of patients can be compared by greater/less, twice/half |
| ISBN numbers for books | Quantitative (Interval) | There is no `0` value as they are unique identification values |
| Ability to pass light in the terms of the following values | Qualitative (Ordinal) | As can only be greater or less than 1 |
| Military rank | Qualitative (Ordinal) | As can only be greater or less than lieutenant |
| Latitude/Longitude values | Quantitative (Interval) | Does not have 0 latitude values |
| Volume of a substance in cubic centimeter | Quantitative (Ratio) | Measurements can have 0 values and can be nominal, ordinal, interval and ratio |
| Coat check number | Quantitative (Interval) | There is no `0` value as they are unique identification values |

**Q2.** (17 points) [Linting Xue] Data Transformation.

(a) Please specify what kind of data transformation needs to be done under the following

situations, and give a brief description about your answer:

i. (5 points) Contrast stretching (often called normalization) is a simple image

enhancement technique that attempts to improve the contrast in an image.

Suppose our images can be represented in a binary format of 16 bits. But

the actual pixel values are distributed between 290 and 3000. Transform (or

stretch) these pixel values into full possible range of the binary values.

ii. (5 points) The domain of variable x is (􀀀1;1), please use a proper function

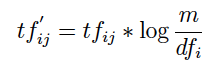
to map x to z 2 (0; 1) (Open set). Make sure that 8z1; z2; z1 > z2 $ (if and

only if) 8x1; x2; x1 > x2

(b) Consider a document-term matrix, where tfij is the frequency of the ith word (term)

in the jth document and m is the number of documents. Consider the variable

transformation that is defined by



where dfi is the number of documents in which the ith term appears, which is

known as the document frequency of the term. This transformation is knows as

the inverse document frequency transformation.

i. (4 points) What is the e\_ect of this transformation is a term occurs in one

document? In every document?

ii. (3 points) What might be the purpose of this transformation?

**Sol.**

**(i)**

Since our Initial Pixel values are n a range of 290 to 3000, we have to perform Normalization for Contrast Stretching them to enhance it to 0 to 65535 range.

Let’s take I as initial dataset, where,

Max = 3000 (Maximum pixel value in the set)

Min = 290 (Minimum pixel value in the set)

Now, to normalize the data,

where, *I= Any random value which we want to normalize*

*NewMax = Maximum intensity value of new Intensity range*

*NewMin = Minimum intensity value of new Intensity range*

*Max = Maximum intensity pixel value of given range*

*Min = Minimum intensity pixel value of given range*

We have,

NewMax = 65,535

NewMin = 0

Max = 3000

Min = 290

Therefore,



For example,

(1) I = 3000

*I’ = 65535*

(2) I = 300



*I’ ≈ 242*

(3) I = 1000



*I’ ≈ 17,170*

In the above problem we learned that Contrast Stretching Normalization is a transformation technique used in Image Processing, in which we enhance or diminish the the pixel intensity of our greyscale image from the given range to desired range. In our problem we have 290 to 3000 as our given pixel range and 0 to 65535 as desired pixel range (As we want to represent in a 16-Bit binary number. So, it’s range is 0-65535). So, we are enhancing the quality of our image by representing it in a 16 bpp image in such a way that every bit of our display is filled by a pixel and making the image more clear.

**(ii)**

We are given,

x = (-∞,∞)

and we have to transform our data values into the range of (0,1)

To perform our transformation we will use Min-Max Scaling to normalize our data.

In Min-Max-Scaling,

where, *xi = value in the range we want to normalize*

*xmin = minimum value in the range*

*xmax = maximum value in the range*

Let’s take an example,

x = (-100,100)

here,

xmin = -100

xmax = 100

But xi can’t be -100 or 100 as it is an open set.

(1) Take x1=50



*z1= 0.75*

(2) Take x2=-90



*z2 = 0.05*

Here, x1>x2 and we get z1>z2. Therefore, condition satisfied and we will get unique values for every value of x.

**(b)**

(i)

1. If a term occurs in only one document then the final transformation will be product of log(total no of document) and the frequency of word in that document.

For example,If a word “Carolina” occurs 4 times in only one document which has 1000 words and out of 100,000 document then after the transformation the weight will be assigned to Carolina in following way:

tfij = 4 / 1000 = 0.004

log (m/dfij) = log(100000/1) = log(105) = 5

Therefore tf’ij will be a product of 5 and 0.004 which is 0.02.

With assessing tf’ij of other words in a document we can check how important the term ‘Carolina’ is to the single document and whether it is part of words which can give idea about the paragraph in few words.

2. If a term occurs in all documents then the final transformation will be zero.

For example,if a term ‘an’ occurs in all the documents then m=dfij ,which will lead to log of 1,therefore the zero weight.

ii)

The main idea behind inverse document frequency transformation is to assess the significance of a word in a single document out of collection of documents.

An intuitive idea behind assessing the significance of the word in a document would be to collect the frequency of the each word in a document and word with most occurrence would be the most significant.

However, the significance depends heavily upon the language and the subject of documents. For example ‘a’,’an’,’the’ are generally most significant words in a document written in an English language ,therefore these words rank higher in the frequency chart but if the same set of words are measured against all the documents then we can deduce that they occur in almost all the documents and therefore they are not significant or rare enough to represent a collective idea behind a document.

These words, though being frequent in a specific document, don’t really separate that document from the others. Hence we have an idf terms which assesses if a word is *rare* (and hence important) while considering all the documents together.

**Q3.** (10 points) [Yihuan Dong]

Describe the advantage and disadvantage of using sampling to reduce the number

of data points.

(a) (2 poin(tbs) (5 points) What is simple random sampling (without replacement)? Ex-

plain when we should use simple random sampling and when we shouldn't use simple

random sampling. Give an example for each situation.

(c) (3 points) Describe what is noise and what is outlier? Which one is acceptable?

**Sol.**

**(a)**

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Advantages** | **Disadvantages** |
| **1** | Smaller data size allows the use of a more expensive algorithm. | There can be a bias or error in collecting data |
| **2** | Low cost for data collection | Difficult to select sample that truly represents the data set. |
| **3** | Having correctly collected a sample, the accuracy of data is high. | Data collected from population may not be accurate. |
| **4** | Using a sample data set takes less time. | Using sampling requires knowledge of statistical analysis and probability. Lack of this knowledge may result in inaccurate. |

**(b)**

Simple random sampling without replacement is the process of randomly selecting an item from data population for the sample and removing it from the population.

Example: We have a stock of certain item and we want to check the quality of the batch. For this, we will use simple random sampling, as it is very difficult for our stock to check each and every item.

Simple random sampling should be used when the data population size is very large, as it will help in faster process.

Eg. Using population census data to find average age where the population size is in billions and it is very difficult to calculate using each value. So, we will use simple random sampling and take fewer values, like a few hundred thousand, for calculations.

Simple random sampling shouldn’t be used when the population is varied as it can fail when to create a proper sample set representing a population.

Eg. When building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample. Using population census data to find state wide population density

**(c)**

Outliers are data objects whose characters are significantly different from rest of the data.

Noise is the modification of original value. It may involve the distortion of a value or the addition of spurious data.

Outliers are legitimate data objects and hence may be of use. Hence outliers are acceptable.

**Q4.** (15 points) [Yuan Zhang] Write code in Matlab, R or Python to perform each of the

following tasks, please report your outputs and key codes in the homework \_le, and also

include your code \_le (end with .m, .r or .py) in the .zip \_le.

(a) (1 point) Generate a 4\*4 matrix A with input from Gaussian Distribution with

mean 5 and variance 1.

(b) (1 point) Access rows 2 and 4 only.

(c) (2 points) Calculate sum of the 3nd row, the diagonal and the 4rd column in the

matrix.

(d) (2 points) Sum of all elements in the matrix (use a for/while loop).

(e) (1 point) Generate a diagonal matrix B with from [2; 3; 4; 5] (using this vector as

the diagonal entries).

(f) (2 points) From A and B, using one matrix operation to get a new matrix C such

that, the \_rst row of C is equal to the \_rst row of A times 2, the second row of C

is equal to the second row of A times 3 and so on.

Fall 2016: CSC 522 CSC 522: HW 1 - Page 3 of 4 8/25/2016

(g) (2 points) From A and B, using one matrix operation to get a new matrix D such

that, the \_rst column of D is equal to the \_rst column of A times 2, the second

column of D is equal to the second column of A times 3 and so on.

(h) (2 points) X = [1; 2; 3; 4]T , Y = [9; 6; 4; 1]T . Computing the covariance matrix of

X and Y in one function, and calculating the result by basic operations (without

using that function).

(i) (2 points) Verifying the equation in X: \_ x2 = (\_x2+\_2(x)), where \_(x) is the estimate

of the standard deviation.

**Sol.**

**Answer a)**

[[ 6.0501036 6.67257808 4.29535052 5.07425193]

[ 5.23760389 5.09027464 6.62905414 4.9271836 ]

[ 3.81108932 4.97893268 5.77593596 5.91148163]

[ 4.4047209 4.79126181 5.95319143 4.32583566]]

**Answer b)**

[[ 5.23760389 5.09027464 6.62905414 4.9271836 ]

[ 4.4047209 4.79126181 5.95319143 4.32583566]]

**Answer c):**

61.9583422586

**Answer d):**

83.9288497778

**Answer e)**

[[2 0 0 0]

[0 3 0 0]

[0 0 4 0]

[0 0 0 5]]

**Answer f)**

[[ 12.10020719 13.34515616 8.59070105 10.14850386]

[ 15.71281166 15.27082391 19.88716242 14.78155079]

[ 15.24435728 19.91573073 23.10374386 23.64592651]

[ 22.0236045 23.95630906 29.76595713 21.62917828]]

**Answer g)**

[[ 12.10020719 20.01773424 17.18140209 25.37125964]

[ 10.47520777 15.27082391 26.51621656 24.63591798]

[ 7.62217864 14.93679805 23.10374386 29.55740814]

[ 8.8094418 14.37378543 23.81276571 21.62917828]]

**Answer h)**

[[ 1.66666667 -4.33333333]

[ -4.33333333 11.33333333]]

**Answer i)**

both are same:

x ̄ 2 = (x̄ 2 +σ 2 (x)) :

7.500000 = 6.250000 + 1.250000

Note: As we have taken random seed function, so, data points may vary after each execution.

**Q5.** (25 points) [Linting Xue] For this exercise, use the provided "locations.csv" \_le, which

contains a list of 1718 data instances. The \_le contains the migration information for

a turkey vulture named 'Tesla' for Fall 2013. (Source: movebank.org) There are 25

columns representing several properties, including the latitude and longitude positions

of Tesla,GPS tag number, date, time, behavior etc. For the purpose of this exercise, we

are concerned with the 'lat' and 'long' columns. Complete the following tasks:

(a) (1 point) Load the \_le and read 'lat' and 'long' columns.

(b) (1 point) Make a 2D plot and label the axes (latitude should be x-axis and longitude

should be y-axis)

(c) Compute the mean of latitude and longitude values. Consider this point as P.

(a) (10 points) Compute the distance between P and the 1718 data points using the following distance measures: Euclidean distance, Mahalanobis distance, City block metric, Minkowski metric (for p=3), Chebyshev distance and Cosine distance.

(b) For each distance measure, identify the 10 points from the dataset that are the closest to the point P.

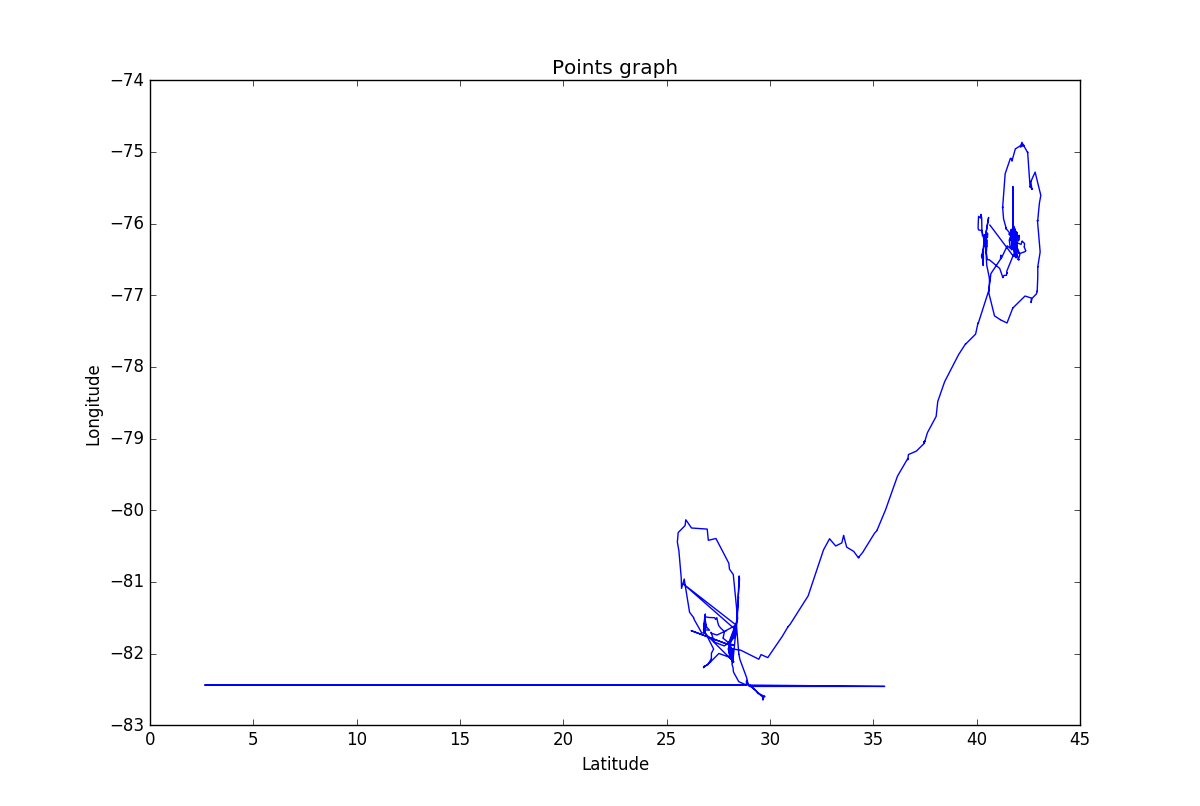
(a) (10 points) Create plots, one for each distance measure. Place an 'X' for P and mark the 10 closest points. To mark them, you could place a circle or draw the line between these closest neighbors and the points 'X'. Make sure the points can be uniquely identified.

(b) (3 points) Verify if the set of points is the same across all the distance measures.

**Sol.**

(a) Executed in script

(b) Figure 1



(c) P = (37.04354412107106, -78.16464476135052)

(c)(b)

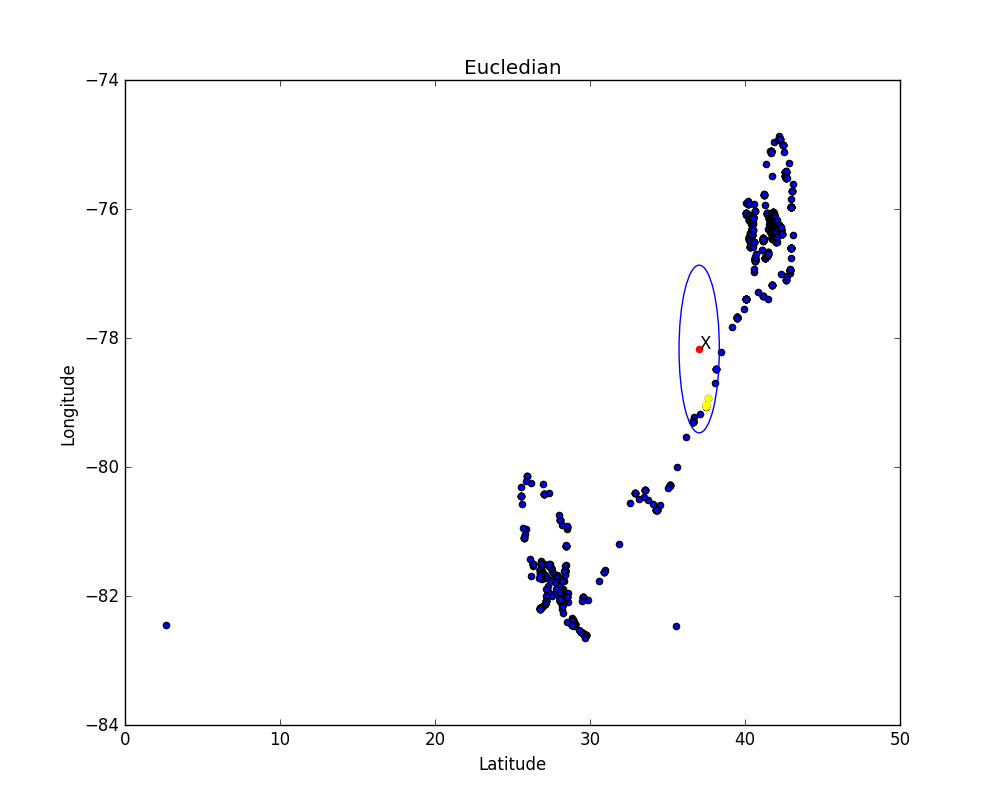
Closest points for each distance measure

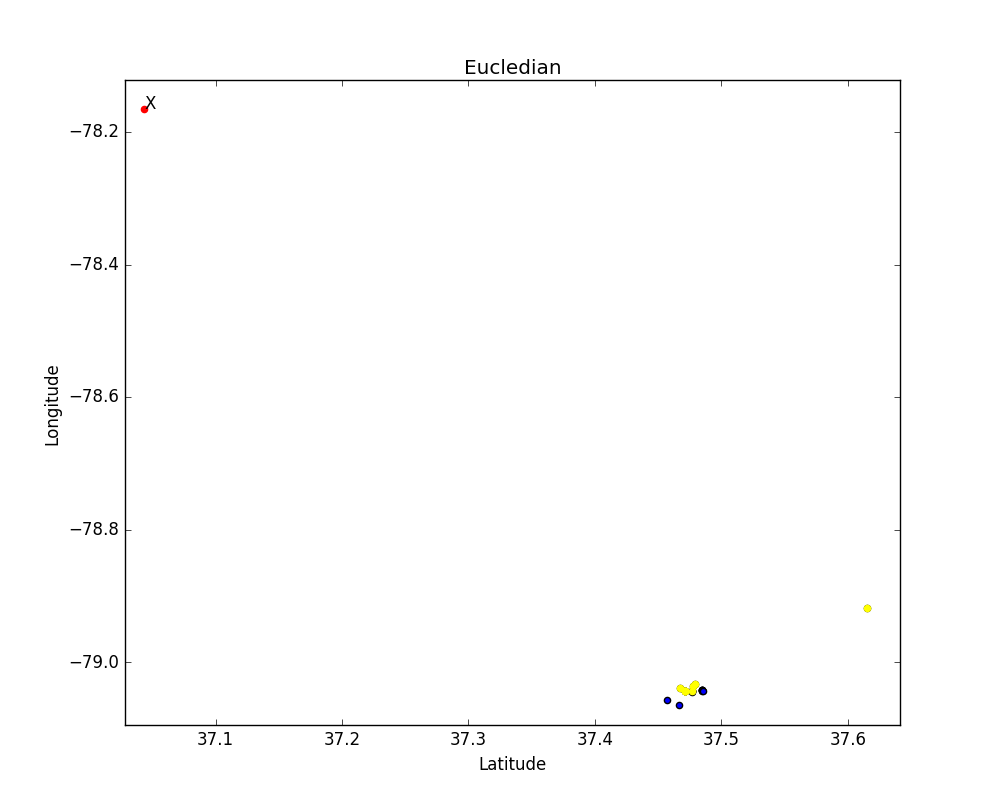
**Table 1**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Distance | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 |
| **Euclidean** | (37.614, -78.917) | (37.479, -79.032) | (37.466, -79.03) | (37.477, -79.036) | (37.471, -79.043) | (37.471, -79.043) | (37.471,-79.044) | (37.472,-79.044) | (37.471,-79.044) | (37.476, -79.043) |
| **Mahalanobis** | (40.623, -76.769) | (40.631,-76.744) | (40.631,-76.745) | (40.625,-76.772) | (40.626,-76.770) | (40.626,-76.772) | (40.626,-76.773) | (40.626,-76.773) | (40.626,-76.773) | (40.626,-76.773) |
| **City block** | (40.638, -76.018) | (40.638,-76.018) | (40.639,-76.018) | (41.770,-76.452) | (41.770,-76.443) | (41.767,-76.437) | (41.770,-76.438) | (41.771,-76.435) | (41.771,-76.436) | (41.772,-76.436) |
| **Minkowski** | (37.614, -78.917) | (37.479, -79.032) | (37.477, -79.036) | (37.466, -79.039) | (37.471, -79.043) | (37.471, -79.043) | (37.471,-79.044) | (37.472,-79.044) | (37.471,-79.044) | (37.476, -79.043) |
| **Chebyshev** | (37.614, -78.917) | (37.479, -79.032) | (37.466, -79.033) | (37.477, -79.036) | (37.471, -79.043) | (37.471, -79.043) | (37.471,-79.044) | (37.472,-79.044) | (37.471,-79.044) | (37.476, -79.043) |
| **Cosine** | (37.466, -79.063) | (37.457, -79.055) | (37.466, -79.039) | (37.471, -79.043) | (37.471, -79.043) | (37.471, -79.043) | (37.471,-79.044) | (37.472,-79.044) | (37.477,-79.044) | (37.477, -79.043) |

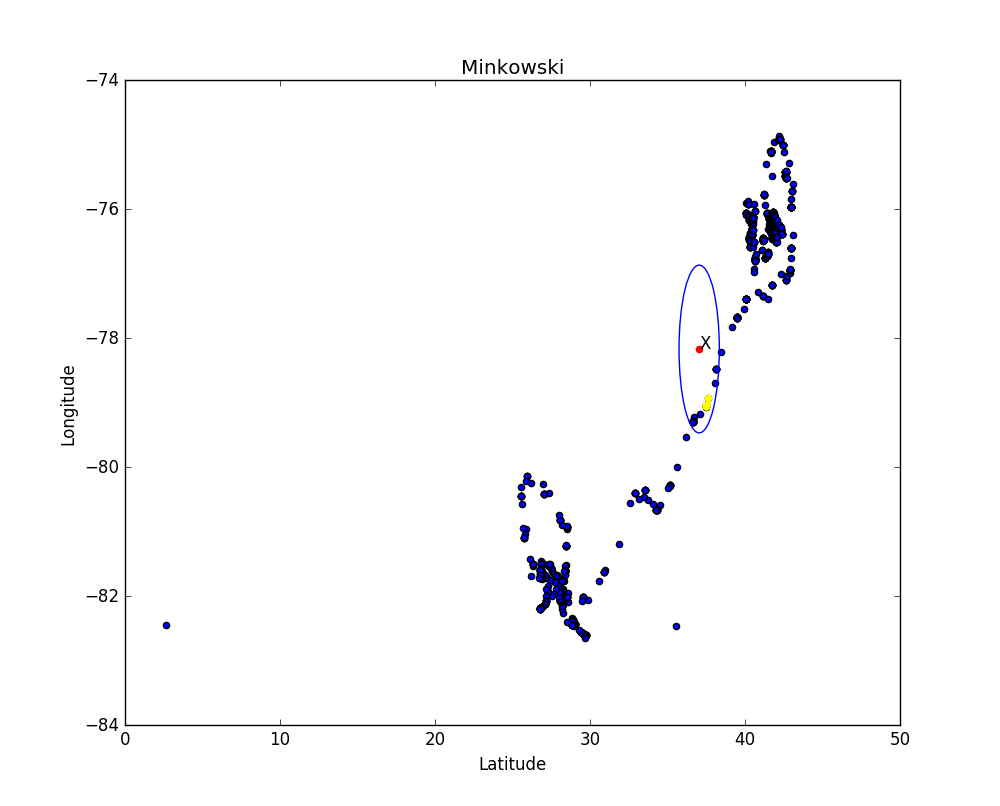
(c)(b)(a) Plots for each distance measure:

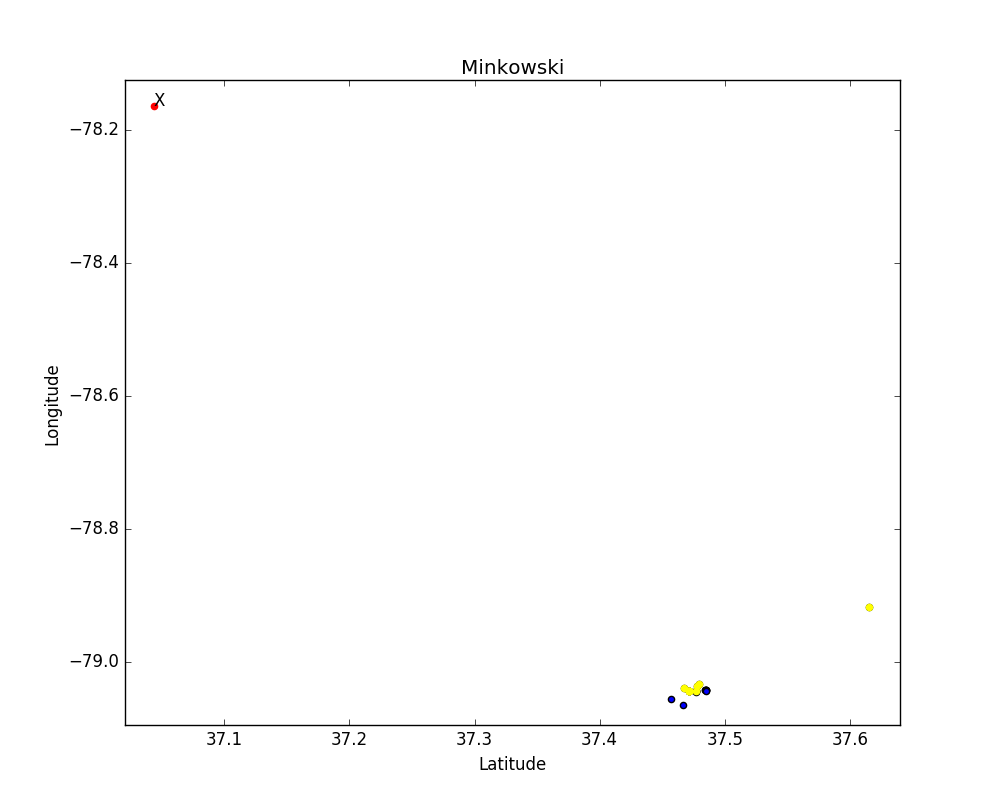
1. Euclidean



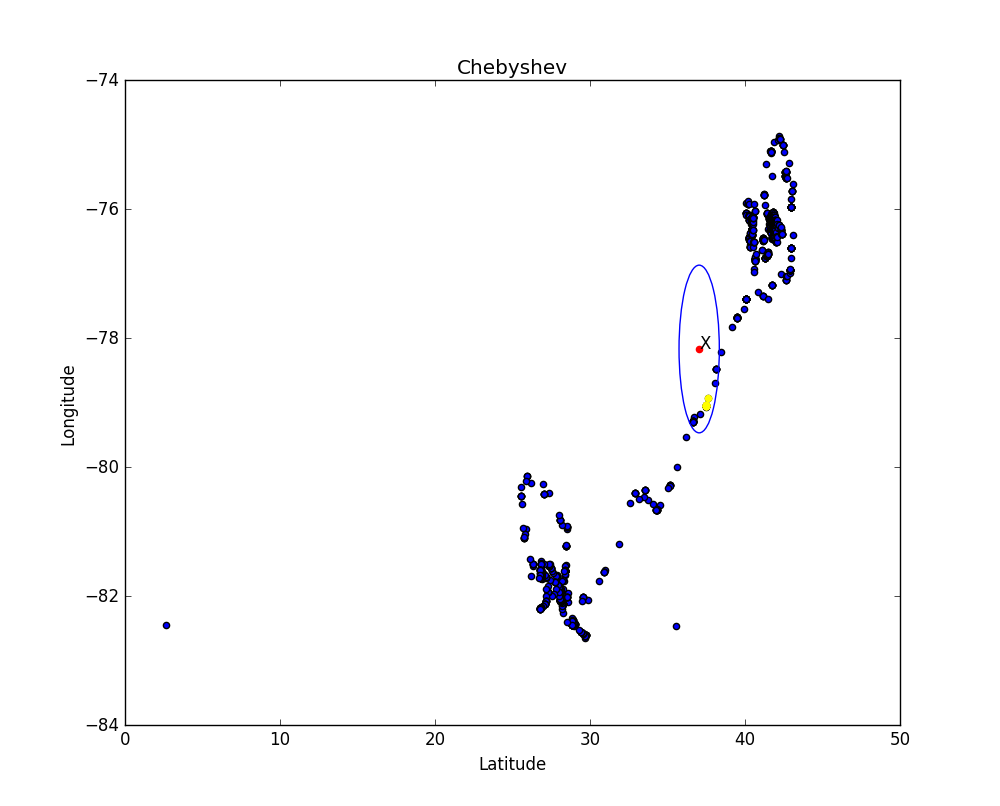


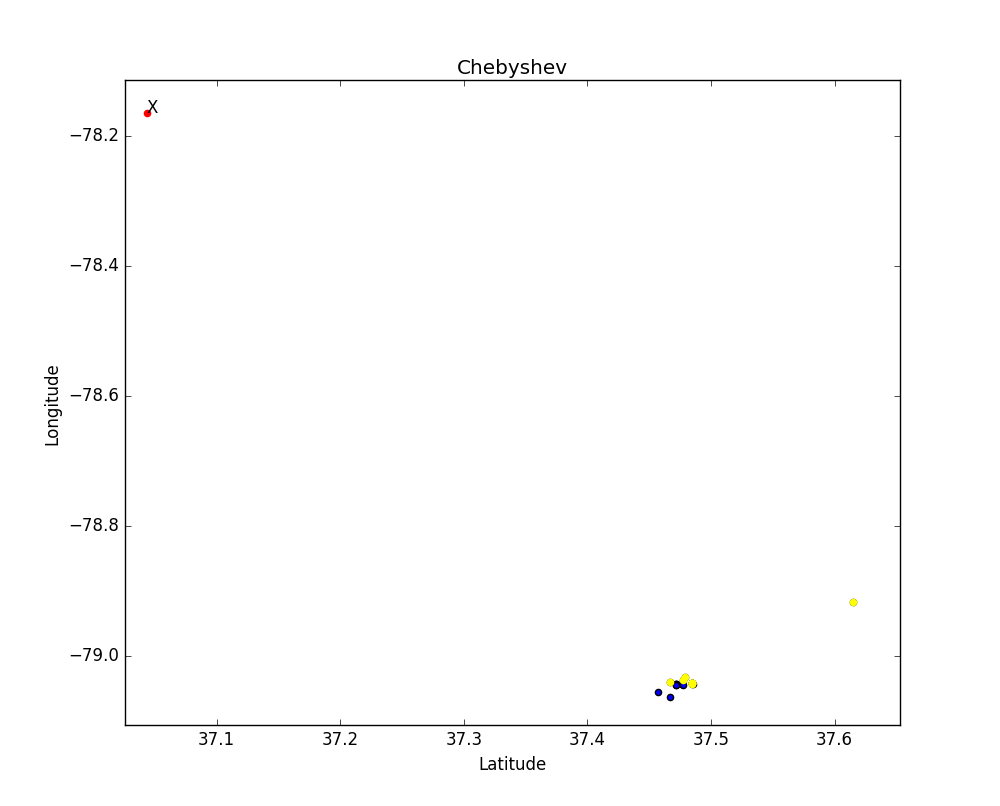
1. Minkowski



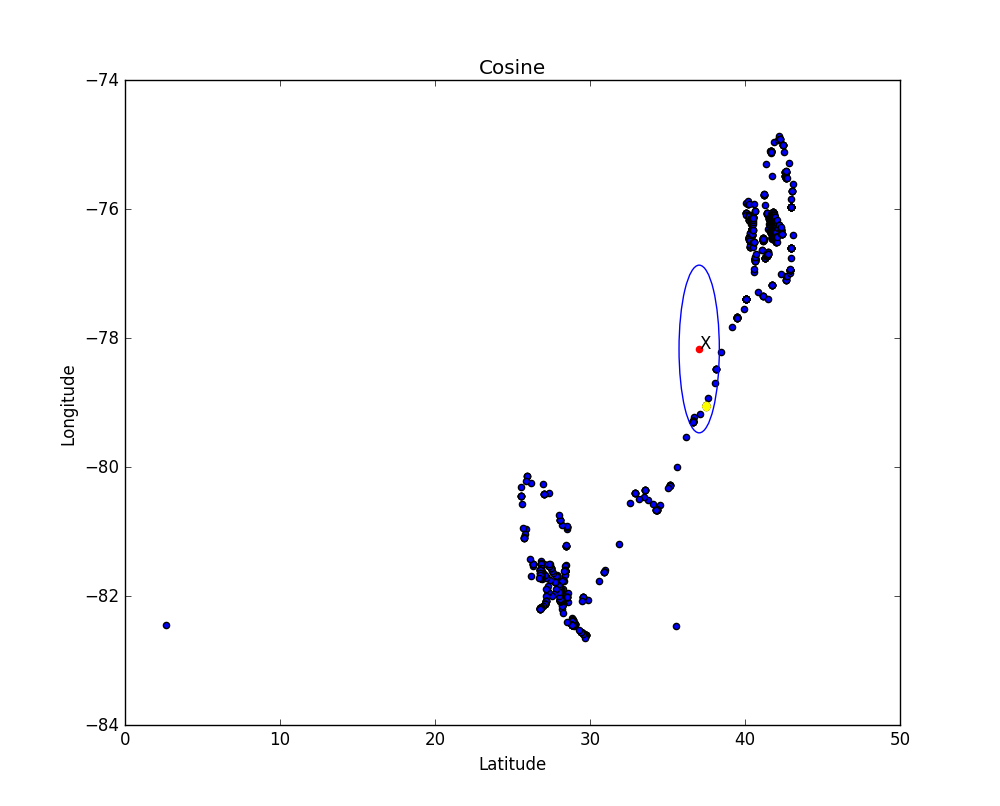


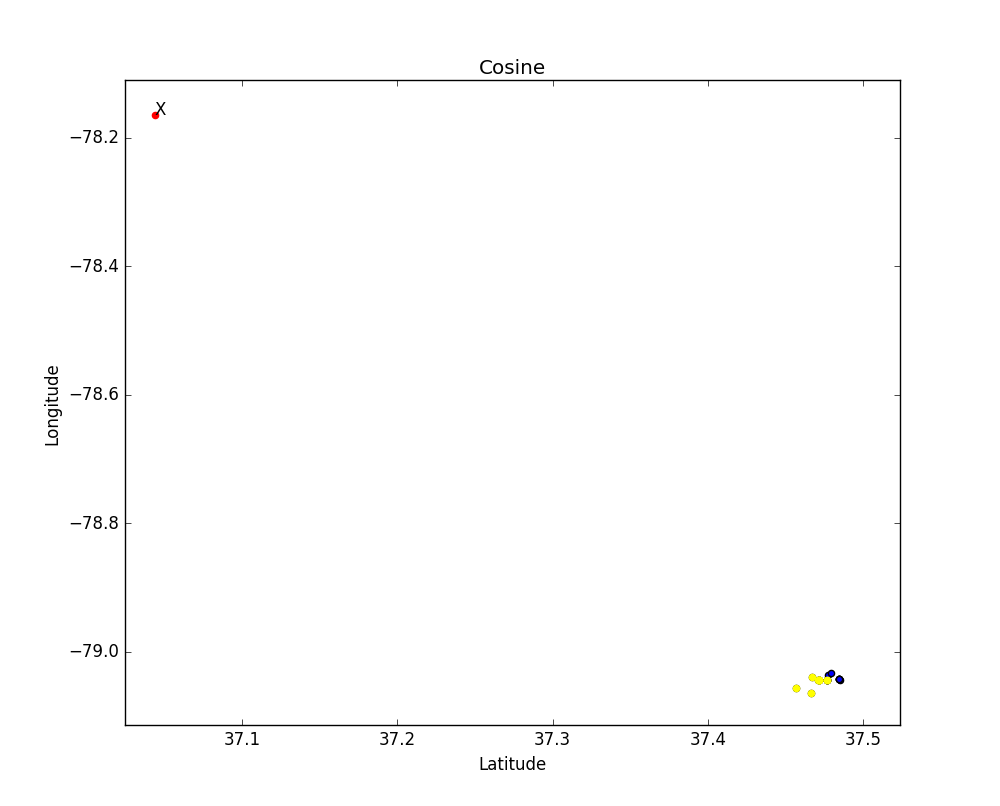
1. Chebyshev



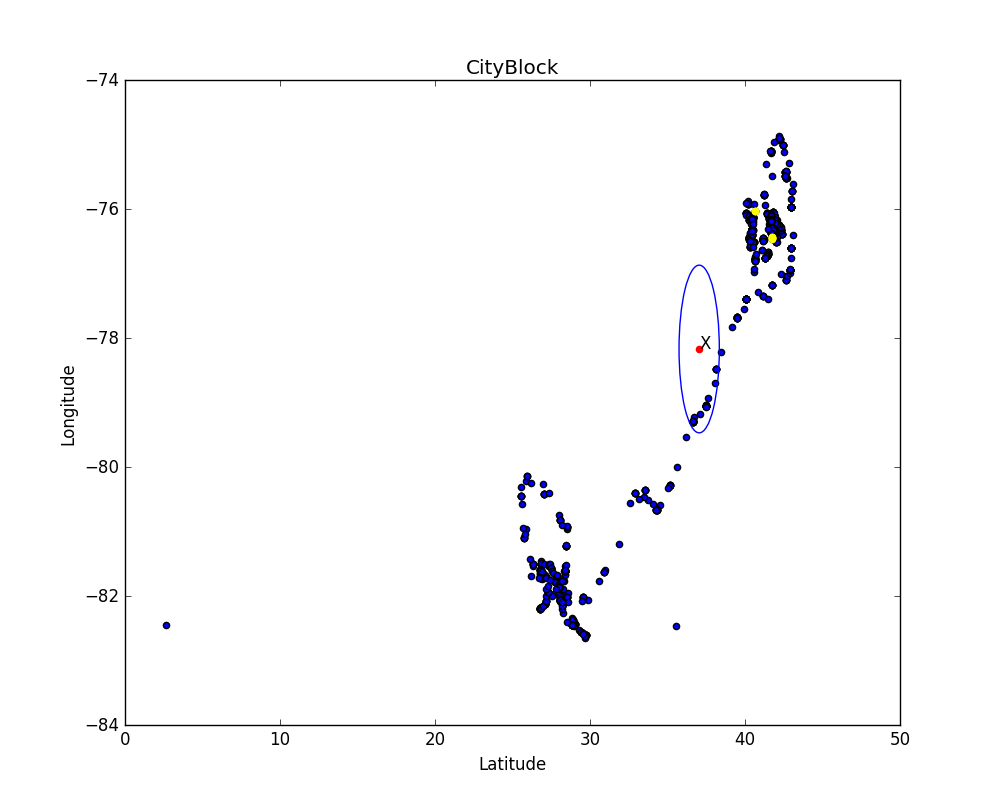


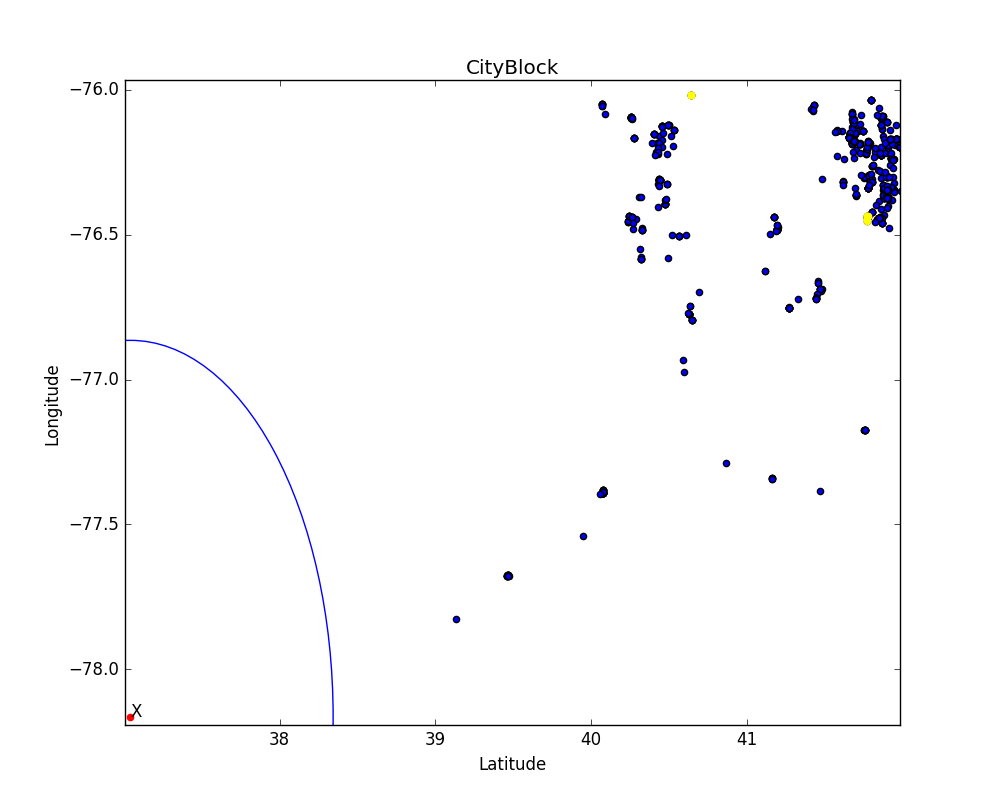
1. Cosine



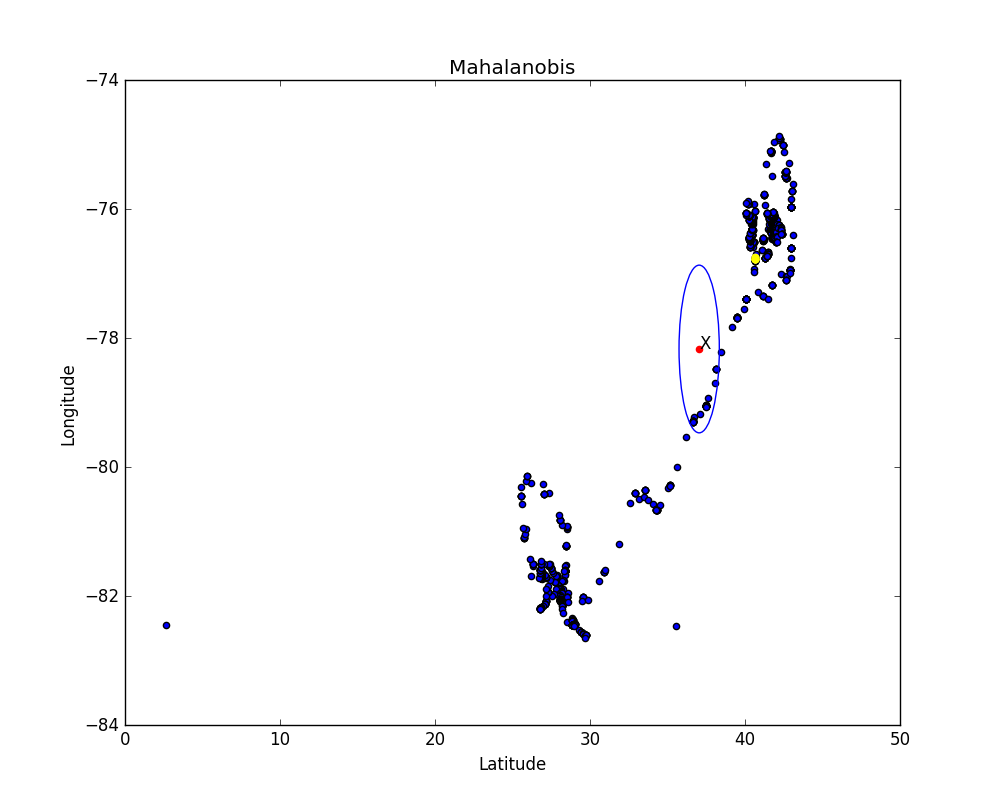


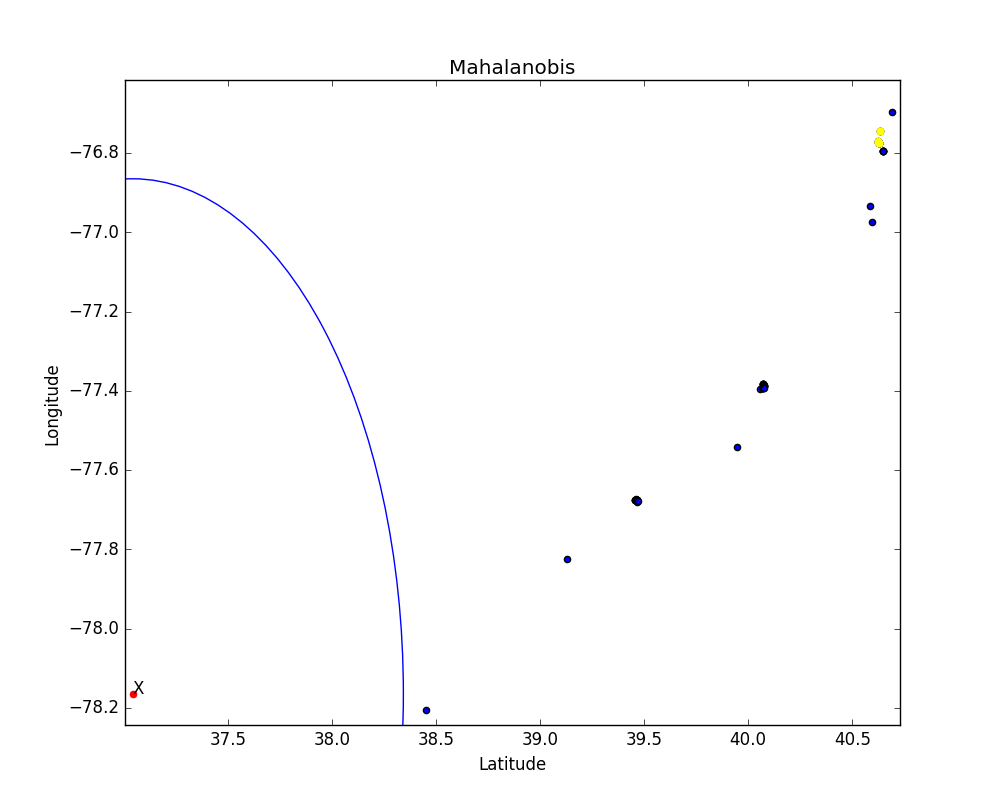
1. City Block





1. Mahalanobis





(c)(b)(b)

Sets of points are same across Euclidean distance, Minkowski metric (for p=3) and Chebyshev distance. Cosine distance also gives similar values but not exactly same as cosine of a variable may vary in calculations. Mahalanobis distance and City Block distance are giving different sets of values for nearest 10 points. For verification, please refer Table 1 of part (c)(b).

**Q6.** (20 points) [Yihuan Dong] In this question, summarize and explore data in the provided \_le hw1q6.csv. This \_le contains 2 columns of 1000 data instances each. Paste in screen shots for each of the questions.

(a) (5 points) Load the data, Compute and report the mean, median, standard deviation, the maximum likelihood of deviation and the range (i.e. the minimum and maximum) for each variable x1, x2.

(b) (2 points) Compute the quantiles for each variable. The quantiles of data set are the 0,25,50,75,and 100 percentiles.

(c) (4 points) Create a histogram for each variable using 10 bins. The scale of the y-axis should be in terms of density. Also, \_tting a density curve to the histogram, In this case, we can simply use normal distribution.

(d) (4 points) Create a quantile-quantile plot (commonly called a QQ plot) for each variable. Include in your plot a line indicating perfect agreement,i.e. y = x. What could this QQ plot be used for? If the data came from a normal distribution, what will happen when we plot the quantiles of our data against the that of a normal distribution?

(e) (5 points) Comment briefly on what you have learned about each variable. Include comparisons between x1 and x2 using location measures (mean, median, etc.), spread measures (standard deviation), and the shape of the histograms. Qualitatively, does either variable appear to have come from a normal distribution? Why?

**Sol.**

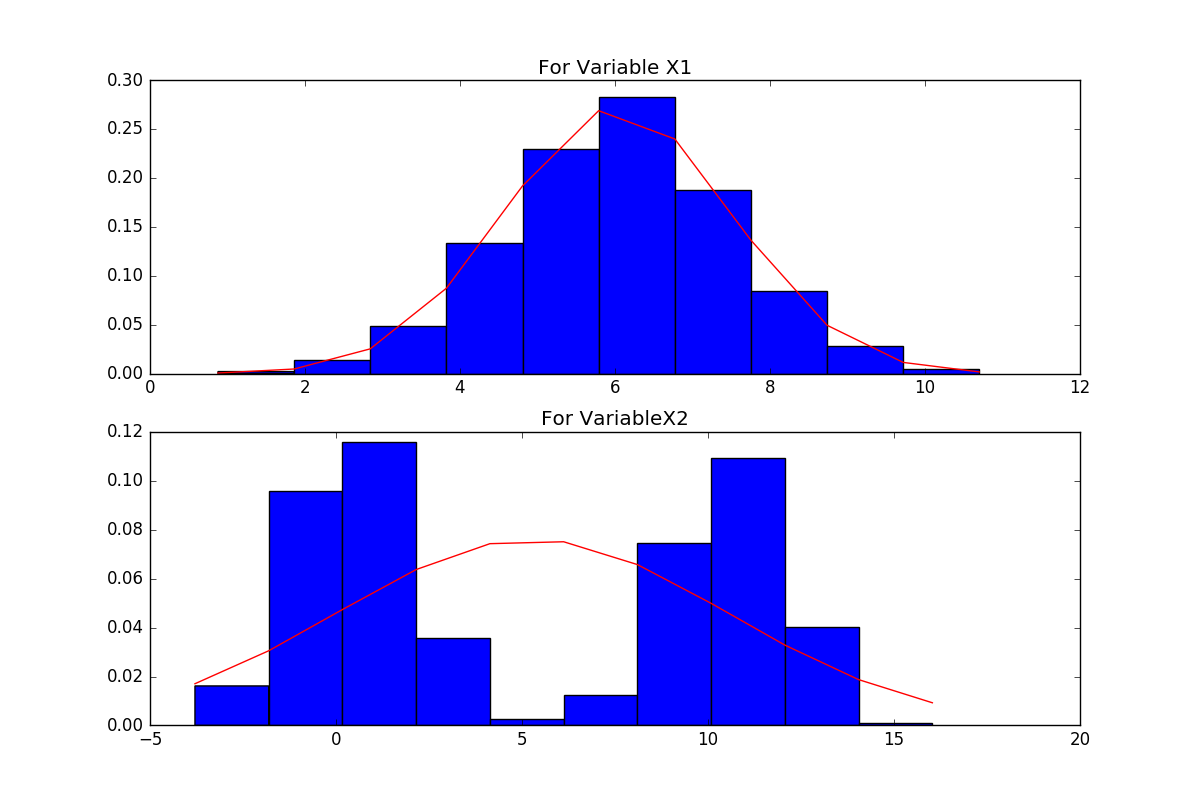
**Answer : (a)**

|  |  |
| --- | --- |
| **x1** | **x2** |
| Mean = 6.03240505 | Mean = 5.273637641 |
| Median = 6.05195 | Median = 3.0669 |
| Standard Deviation = 1.46280875222 | Standard Deviation = 5.23952944429 |
| Range = 9.82405 | Range = 19.8225 |

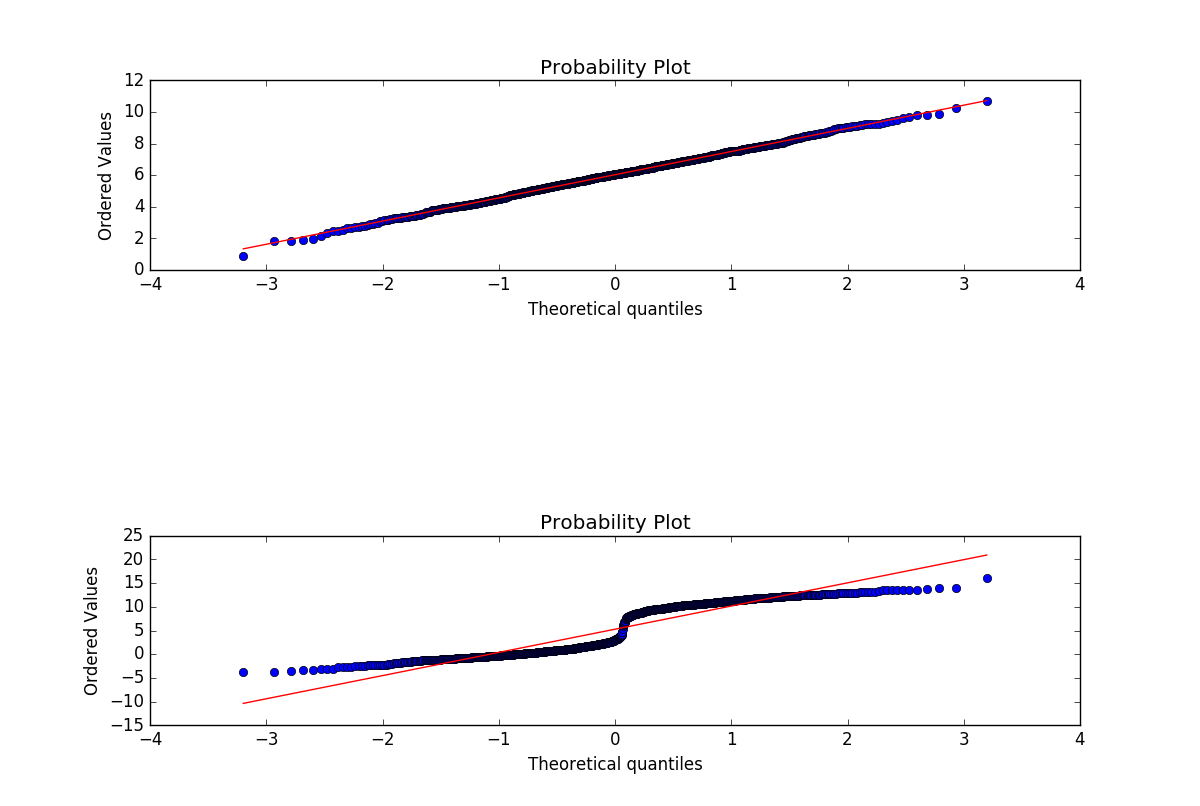
**Answer : (b)**

|  |  |
| --- | --- |
| **x1** | **x2** |
| Q1 = 0.87695 | Q1 = -3.7915 |
| Q2 = 5.082675 | Q2 = 0.367945 |
| Q3 = 6.05195 | Q3 = 3.0669 |
| Q4 = 6.990675 | Q4 = 10.457 |
| Q5 = 10.701 | Q5 = 16.031 |

**Answer : (c)**



**Answer : (d)**



The q-q plot or Quartile-Quartile plot is used to check whether the given data set is normalized or not. If the data comes from a normal distribution then the data points will be overlapping the 45-degree line otherwise they will be scattered away from that line.

**Answer : (e)**

The mean value of 1st variable is approximately same as medium, which means that the distribution is approximately symmetrical.

The mean value of 2nd variable is not close to medium, mean > median, which means that the distribution is negatively skewed.

The standard deviation of 1st variable is 1.46, which means the data is not skewed but the standard deviation of 2nd variable is 5.23 which means data is skewed.

The histogram shape of the first variable shows that data is evenly distributed about the mean that means it is normalized and shape of second variable shows that data is not evenly distributed about the mean and is not normalized.

The first variable values comes from a normalized data as the QQ-plot of that data is more evenly distributed along the line vector as compared to 2nd variable where the data points are more scattered from the line vector.