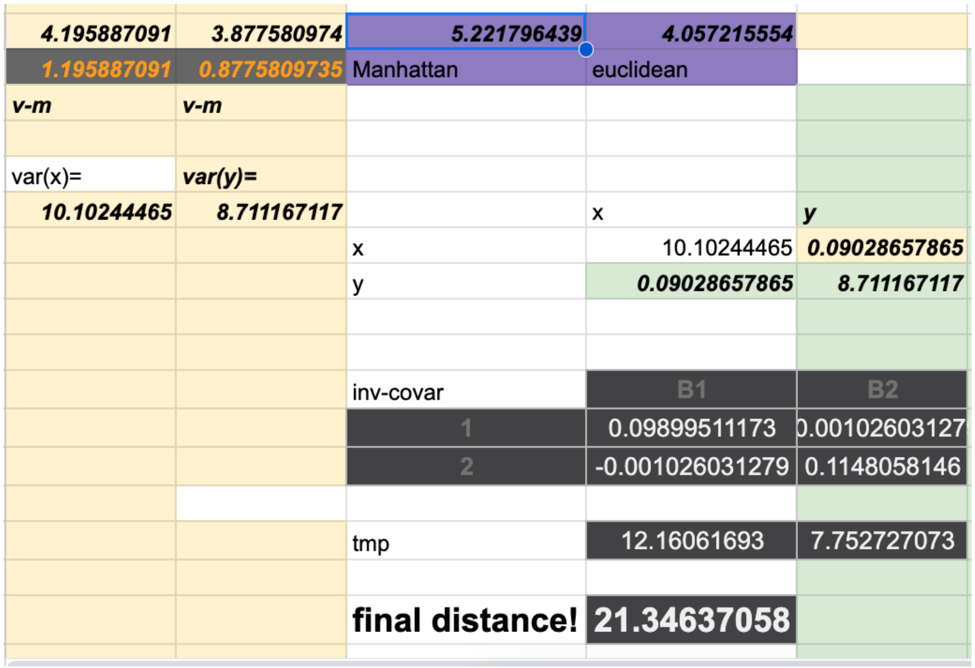
Carl Cortez
CIS 735
HW 1

***1. a. (35 pts) Given the data in two classes of data: dataSet1a.csv (Class a) and dataSet1b.csv (Class b), classify the point (3,3) as being in either Class a or Class b using the Manhattan distance, Euclidean distance, and Mahalanobis distance. For the Minkowski distances measure use the centroid (mean) of the data sets as the exemplar of the data.***

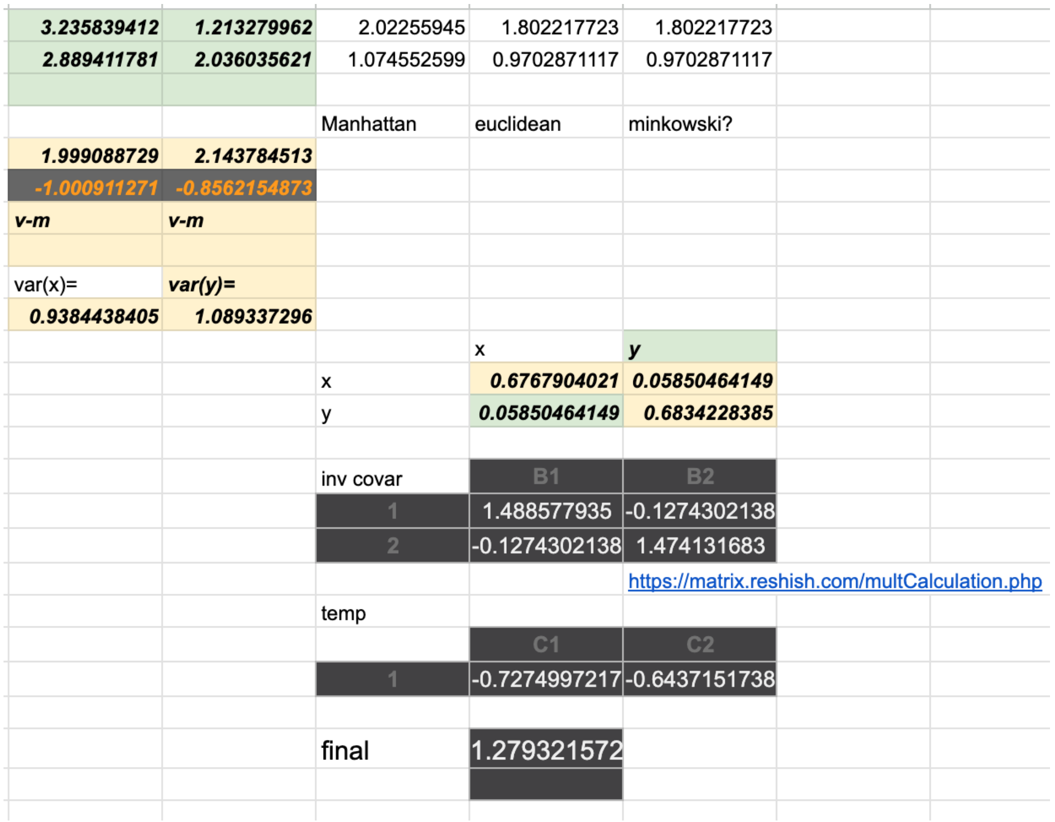
1a) google sheet with work.

Data Set 1



| | | | | |
|---|---|---|---|---|
| ***4.195887091*** | ***3.877580974*** | ***5.221796439*** | ***4.057215554*** | |
| ***1.195887091*** | ***0.8775809735*** | Manhattan | euclidean | |
| ***v-m*** | ***v-m*** | | | |
| | | | | |
| var(x)= | ***var(y)=*** | | | |
| ***10.10244465*** | ***8.711167117*** | | x | ***y*** |
| | | x | 10.10244465 | ***0.09028657865*** |
| | | y | ***0.09028657865*** | ***8.711167117*** |
| | | | | |
| | | | | |
| | | inv-covar | B1 | B2 |
| | | 1 | 0.09899511173 | 0.00102603127 |
| | | 2 | -0.001026031279 | 0.1148058146 |
| | | | | |
| | | tmp | 12.16061693 | 7.752727073 |
| | | | | |
| | | **final distance!** | **21.34637058** | |

Data set 2



| | | | | |
|---|---|---|---|---|
| ***3.235839412*** | ***1.213279962*** | 2.02255945 | 1.802217723 | 1.802217723 |
| ***2.889411781*** | ***2.036035621*** | 1.074552599 | 0.9702871117 | 0.9702871117 |
| | | | | |
| | | Manhattan | euclidean | minkowski? |
| ***1.999088729*** | ***2.143784513*** | | | |
| ***-1.000911271*** | ***-0.8562154873*** | | | |
| ***v-m*** | ***v-m*** | | | |
| | | | | |
| var(x)= | ***var(y)=*** | | | |
| ***0.9384438405*** | ***1.089337296*** | | | |
| | | | x | ***y*** |
| | | x | ***0.6767904021*** | ***0.05850464149*** |
| | | y | ***0.05850464149*** | ***0.6834228385*** |
| | | | | |
| | | inv covar | B1 | B2 |
| | | 1 | 1.488577935 | -0.1274302138 |
| | | 2 | -0.1274302138 | 1.474131683 |
| | | | | https://matrix.reshish.com/multCalculation.php |
| | | temp | | |
| | | | C1 | C2 |
| | | 1 | -0.7274997217 | -0.6437151738 |
| | | | | |
| | | final | 1.279321572 | |

***1. b. (15 pts) When you vary the p (exponential) values e.g. 0.5, 1.5, 100, how does that affect the Minkowski distance values (use examples)?***

For the Euclidean distance, the power of the difference matches the root around the entire summation.

Example: p=100

$$\sqrt[100]{\sum_{i=1}^{n}(x_i - y_i)^{100}}$$

p=.5

$$\sqrt[.5]{\sum_{i=1}^{n}(x_i - y_i)^{.5}}$$

p=1.5

$$\sqrt[1.5]{\sum_{i=1}^{n}(x_i - y_i)^{1.5}}$$

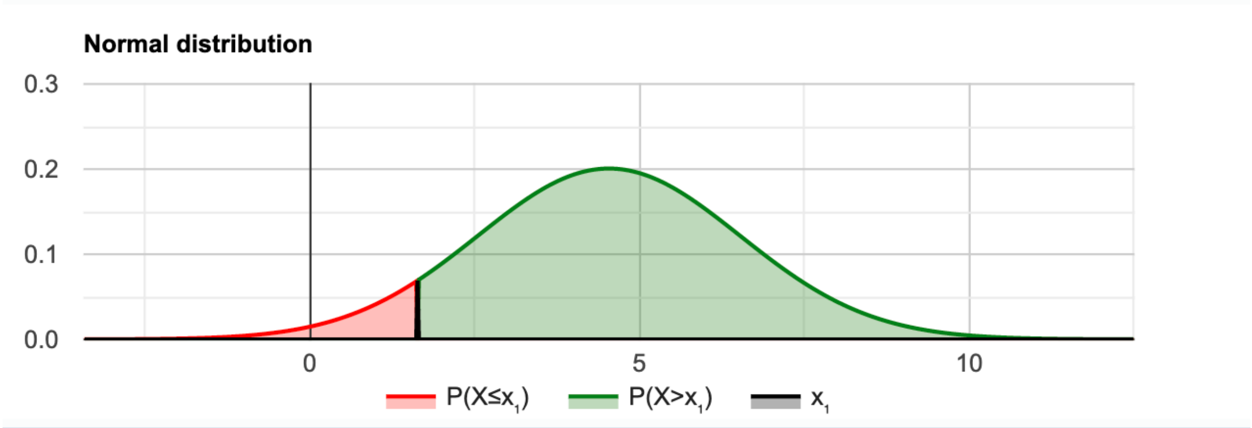
***2. (50 pts) Read in the dataSet2.csv values, for parametric density estimation assume a Gaussian distribution and find the mean and standard deviation of the data. Also, using a histogram as surrogate for nonparametric density estimation.***

Read in values found in this google sheet.

| *Mean* | *SD* |
|---|---|
| *4.538553521* | *1.98789804* |

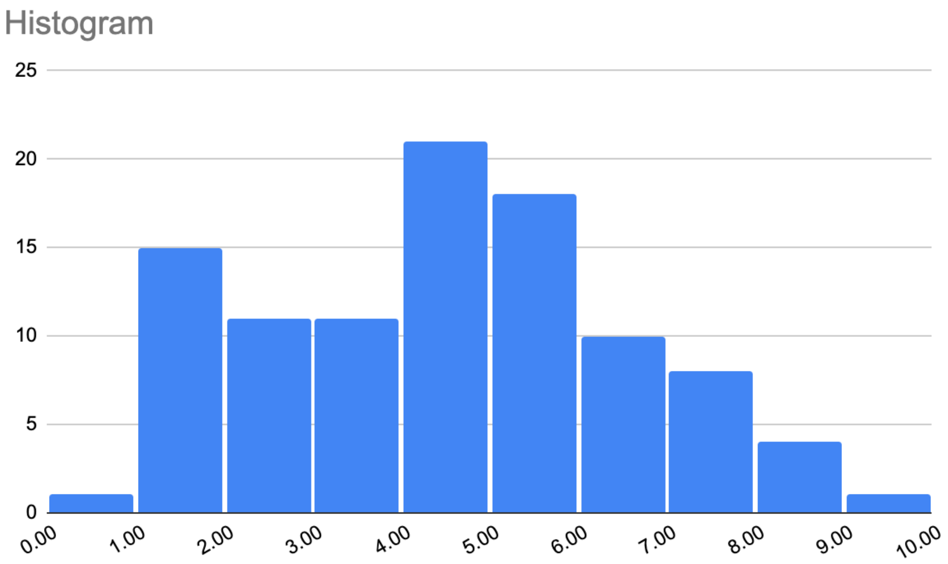
***For the parametric assumption, what is the mode/modes of the data?***

Around 4ish



***For a nonparametric assumption, what is the mode/modes of the data?***

Between 4 and 5.



***<u>What can you assume about the data using the nonparametric density estimate that you could not using a parametric density estimation?</u>***

The quantity of results in the data that are in each bin.