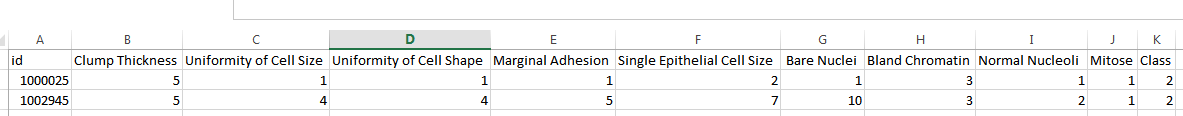
# LAB 9

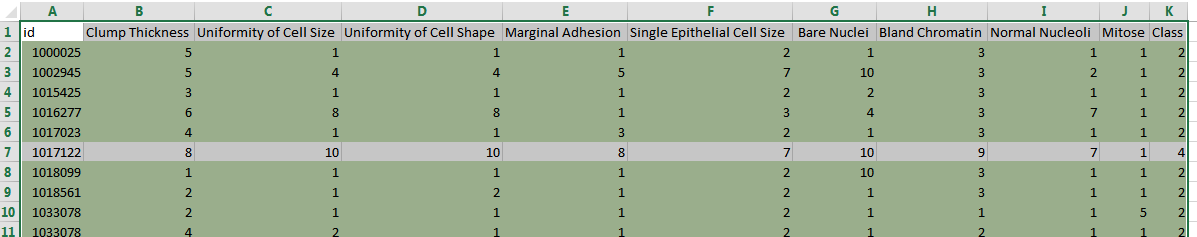
Excel PivotTable and PivotChart provide ways to analyze large datasets. In this lab, we will import a real world breast cancer data into excel and try to gain a rudimentary understanding of how raw data can be transformed into powerful diagnostic/prediction tools using statistically based algorithms and computers/machines.

*Please refer to*[*Intro to Database lecture*](https://www.google.com/url?q=https://docs.google.com/presentation/d/12hoJip_2rBBHu4_48q9CHxgpUeD_lnptv0u9ewr5W68/pub?start%3Dfalse%26loop%3Dfalse%26delayms%3D3000&sa=D&ust=1490816144417000&usg=AFQjCNGGQ-Dhr2s9YfPQfh_YrIe4J_Gf3w)*/in class demo in order to complete these activities. Please send workbook to*[*shconley@syr.edu*](mailto:schonley@syr.edu)*once steps have been completed.*

1. Create a Table in Excel containing the data. The attribute or column names are located [here](https://www.google.com/url?q=http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names&sa=D&ust=1490816144419000&usg=AFQjCNGkyC1yrL2ba7NicsjhRjkhEkJ3xw). The tuples or row are located [here](https://www.google.com/url?q=http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data&sa=D&ust=1490816144421000&usg=AFQjCNFAh93VyJEN12xZnlpUpSSj4HQfOg).

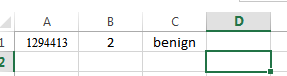


1. Highlight all rows where the patient does not have cancerous cells in a color of your choosing.



**I have used condition formatting and formula of =INDIRECT("K"&ROW())=2**

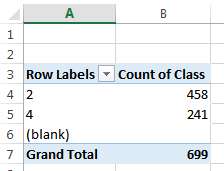
1. Use the LookUp formula to determine if a Patient with ID # 1294413 has cancerous cells/malignant tumor.



**B1 cell formula is =VLOOKUP(A1,data!A1:K700,11)**

**C1 cell formula is =IF(B1=2,"benign","malignant")**

1. Create a PivotTable with two Rows. One row will represent patients with benign cells (Class 2 patients)  and the Class 4 patients with malignant cells.



1. Create values that contain the average for each of the cell lab analysis types (Clump Thickness, Uniformity of Size, etc.). Note you should have eight columns in your pivot table and two rows. Make sure to remove the grand total row and ensure that there are no decimal places/points in the values.



1. Use PivotCharts to visualize the table and add error bars and data labels to the columns.
2. Choose a column besides BareNuclei and determine the standard deviation for each patient class along with the max and min values for each by adding a new columns to your PivotTable. Based on the standard deviation, min, max, and average difference between patient classes, what range would you use to determine if the patient has breast cancer? How many patients in this dataset would receive a false negative diagnosis? (They have malignant cells, but would be told they do not based on the range you just specified?)  Please use an excel formula at the bottom of the original data table and provide the percentage.

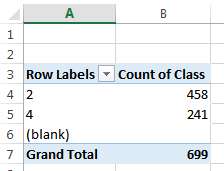
#### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row Labels** | **Max of Bare Nuclei** | **Min of Bare Nuclei** | **StdDev of Bare Nuclei** | **Average of Bare Nuclei** |
| 2 | 10 | 1 | 1.18 | 1.35 |
| 4 | 10 | 1 | 3.12 | 7.63 |

Patients with breast cancer have Bare Nuclei in the range of Average+/-standard deviation.

So, (7.63-3.12,7.63+3.12)= (4.51; 10.75)

=COUNTIF(G2:G700,">4.51") yields 204 patients, but in reality there 241 patients with malignant class. So 241–204=37 people would have diagnosed wrongly, i.e. they have cancer but they are told that they don’t have cancer. Percentage of error would be 37/699=5.3%



|  |
| --- |
| predicted cancer |
| 204 |
| error rate |
| 5.3% |