**Data Wrangling**

We will explore a breast cancer data set and use pandas to clean, reshape, massage and give us a clean data set, all of this will help dramatically increase the quality of our data.

The following functionality were done for cleaning and filtering records:

1. Data exploration — columns, unique values in a column, describe, duplicates
2. Dealing with missing values — quantifying missing values per column, filling & dropping missing values
3. Filtering data
4. Reshaping data — pivot tables, joins, grouping and aggregating
5. Other — Making descriptive columns, element-wise conditional operations

**Data exploration**

After reading the data set into pandas and displaying the column names along with their data types.

patient\_id int64

clump\_thickness float64

cell\_size\_uniformity float64

cell\_shape\_uniformity int64

marginal\_adhesion int64

single\_ep\_cell\_size int64

bare\_nuclei object

bland\_chromatin float64

normal\_nucleoli float64

mitoses int64

class object

doctor\_name object

Based on this, we can assume that patient\_id is a unique identifier, class is going to tell us whether the tumor is malignant (cancerous) or benign (not cancerous). The remaining columns are numeric medical descriptions of the tumor, except for the doctor\_name which is a categorical feature.

**Dealing with missing values**

With every data set it is vital to evaluate the missing values. How many are there? Is it an error? Are there too many missing values? Does a missing value have a meaning relative to its context?

We can sum up the total missing values using the following:

Now that we have identified our missing values, we have a few options. We can fill them in with a certain value (zero, mean/max/median by column, string) or drop them by row. Since there are few missing values, we can drop the rows to avoid skewing the data in further analysis.

**Inspecting duplicates**

To view repeating rows we can start off by looking at the number of unique values in each column.

We saw there are 699 rows, but there are only 645 unique patient\_id’s. This could mean that some patient appear more than once in the data set, so i isolated the patients.

**Filtering data**

We need to remove patients that show up more that 2 times in the data set and hence was removed.

**Reshaping data**

The data set has elements of categorical data in the “doctor\_name” column. To feed this data into a machine learning pipeline, we will need to convert it into an encoded column. This was done using a sci-kit learn package or through pandas to demonstrate the pivoting and merging functionality.

**Row-wise Operations**

Renamed elements within a column based on its value and created a new column that yields a specific value based on multiple attributes within the row.

For this example created a new column that categorizes patients cell as normal or abnormal based on its attributes.

**Conclusion**

Although some of these data manipulation steps can be done in SAS and excel. Doing it in python not only allowed to connect the data to vast open source resources in computer vision, machine and deep learning, but also for ETL automation purposes and more.

**Understanding**

Data Wrangling is the process of converting and mapping data from its raw form to another format with the purpose of making it more valuable and appropriate for advance tasks such as Data Analytics and Machine Learning.