Cheat Sheet: Data Wrangling

Package/ Method	Description	Code Example
Replace missing data with frequency	Replace the missing values of the data set attribute with the mode common occurring entry in the column.	1 2 • `MostFrequentEntry = df['attribute_name'].value_counts().idxmax()` • `df['attribute_name'].replace(np.nan,MostFrequentEntry,>df['attribute_name'].replace(np.nan,MostFrequentEntry, inplace=True)` Copied!
Replace missing data with mean	Replace the missing values of the data set attribute with the mean of all the entries in the column.	1 2 • `AverageValue=df['attribute_name'].astype(<data_type>).mean(axis=0)` • `df['attribute_name'].replace(np.nan, AverageValue, inplace=True)` Copied!

Fix the data types	Fix the data types of the columns in the dataframe.	1 2 3 • df[['attribute1_name', 'attribute2_name', ...]] = • df[['attribute1_name', 'attribute2_name', ...]].astype('data_type') • #data_type is int, float, char, etc. Copied!
Data Normalization	Normalize the data in a column such that the values are restricted between 0 and 1.	1 • df['attribute_name'] = df['attribute_name']/df['attribute_name'].max() Copied!
Binning	Create bins of data for better analysis and visualization.	1 2 3 4 5 6 • bins = np.linspace(min(df['attribute_name']), max(df['attribute_name'],n) • # n is the number of bins needed

		GroupNames = ['Group1','Group2','Group3,...] df['binned_attribute_name'] = pd.cut(df['attribute_name'], bins, labels=GroupNames, include_lowest=True) Copied!
Change column name	Change the label name of a dataframe column.	1 df.rename(columns={'old_name':\'new_name'}, inplace=True) Copied!
Indicator Variables	Create indicator variables for categorical data.	1 2 dummy_variable = pd.get_dummies(df['attribute_name']) df = pd.concat([df, dummy_variable],axis = 1) Copied!