Data Analysis with Python

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.	```python\nMostFrequentEntry = df['attribute_name'].value_counts().idxmax()\ndf['attribute_name'].replace(np.nan,MostFrequentEntry,>df['attribute_name'].replace(np.nan,MostFrequentEntry, inplace=True)\n```
Replace missing data with mean	Replace the missing values of the data set attribute with the mean of all the entries in the column.	```python\nAverageValue=df['attribute_name'].astype(<data_type>).mean(axis=0)\ndf['attribute_name'].replace(np.nan, AverageValue, inplace=True)\n```
Fix the data types	Fix the data types of the columns in the dataframe.	```python\ndf[['attribute1_name', 'attribute2_name', ...]] =\ndf[['attribute1_name', 'attribute2_name', ...]].astype('data_type')\n#data_type is int, float, char, etc.\n```
Data Normalization	Normalize the data in a column such that the values are restricted between 0 and 1.	```python\ndf['attribute_name'] =\ndf['attribute_name']/df['attribute_name'].max()\n```
Binning	Create bins of data for better analysis and visualization.	```python\nbins = np.linspace(min(df['attribute_name']),\nmax(df['attribute_name'],n)\n# n is the number of bins needed\nGroupNames = ['Group1','Group2','Group3',...]\ndf['binned_attribute_name'] =\npd.cut(df['attribute_name'], bins, labels=GroupNames, include_lowest=True)\n```
Change column name	Change the label name of a dataframe column.	```python\ndf.rename(columns={'old_name':\'new_name'}, inplace=True)\n```
Indicator Variables	Create indicator variables for categorical data.	```python\ndummy_variable = pd.get_dummies(df['attribute_name'])\ndf = pd.concat([df, dummy_variable],axis = 1)\n```